Two Causal Principles for Improving Visual Dialog:
Visual Dialog Challenge 2019 Winner Report

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 is a winner report from team MReaL-BDAI for Visual Dialog Challenge 2019. We present ***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 Visual Dialog 2019 Challenge 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 the 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.*

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 give a good answer A. Our community has always considered VQA [4] 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 for VQA [24, 37], as well as for natural language processing [38], 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% [12].

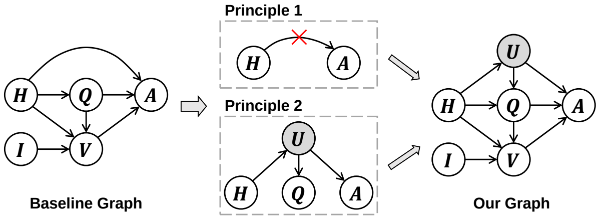


Figure 1. Causal graphs of VisDial models (baseline and ours). H: dialog history. I: image. Q: question. V: visual knowledge. A: answer. U: user preference. Shaded U denotes unobserved confounder. See Section 3.2 for detailed definitions.

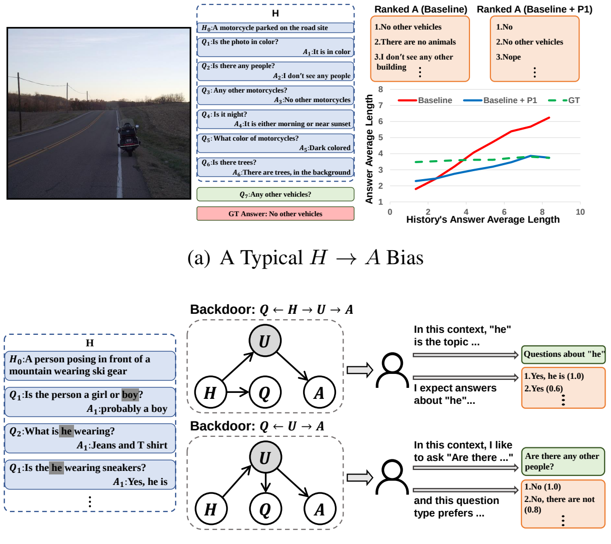
However, in this paper, we want to highlight an important fact: *VisDial is essentially NOT VQA!* And this fact is so profound that all the common heuristics in the vision-language community — such as the fusion tricks [37, 44] and attention variants [24, 26] — cannot appreciate the difference. Instead, we introduce the use of *causal inference* [28, 29]: 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 VisDial models[1] in literature [9, 23, 39, 27], promoting them to the state-of-the-art in Visual Dialog 2019 Challenge [2].

Principle 1 *(P1): Delete the 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.


H
H_0:A motorcycle parked on the road site
Q_1:Is the photo in color?
A_1:It is in color
Q_2:Is there any people?
A_2:I don't see any people
Q_3: Any other motorcycles?
A_3:No other motorcycles
Q_4: Is it night?
A_4:It is either morning or near sunset
Q_5: What color of motorcycles?
A_5:Dark colored
Q_6:Is there trees?
A_6:There are trees, in the background
Q_7:Any other vehicles?
GT Answer: No other vehicles
Ranked A (Baseline)
1.No other vehicles
2.There are no animals
3.I don't see any other building
Ranked A (Baseline + P1)
1.No
2.No other vehicles
3.Nope
Baseline
Baseline + P1
GT
Answer Average Length
1
2
3
4
5
6
7
8
0
2
4
6
8
10
History's Answer Average Length
Backdoor: $Q \leftarrow H \rightarrow U \rightarrow A$
H
H_0:A person posing in front of a mountain wearing ski gear
Q_1:Is the person a girl or boy?
A_1:probably a boy
Q_2:What is he wearing?
A_1:Jeans and T shirt
Q_1:Is the he wearing sneakers?
A_1:Yes, he is
U
H
Q
A
In this context, "he" is the topic ...
Questions about "he"
I expect answers about "he"...
1.Yes, he is (1.0)
2.Yes (0.6)
Backdoor: $Q \leftarrow U \rightarrow A$
U
H
Q
A
In this context, I like to ask "Are there ..."
Are there any other people?
and this question type prefers ...
1.No (1.0)
2.No, there are not (0.8)

(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 data flows. For example, $V \rightarrow A$ and $Q \rightarrow A$ indicate that the visual knowledge V, *e.g.*, the encoded feature from a multi-model encoder, works with the question Q to "dictate" the answer A.

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, 16, 39, 27, 41, 18, 12, 33], 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 annotators of the VisDial dataset [9] were not allowed to copy from the previous Q&A, *i.e.*, $H \nrightarrow A$, and were encouraged to ask consecutive questions that includes co-referenced pronouns like "it" and "those", *i.e.*, $H \rightarrow Q$, and the answer A should be based only on question Q and the reasoned visual knowledge V. Therefore, a good VisDial model is expected to 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 whose lengths are close 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 based only on the association among the sample (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 [29]), which is an indirect causal path from the 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, as the VisDial dataset essentially involves humans in the loop. 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* [29] $P\left(A|do(I, H, Q)\right)$, 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 Visual Dialog Challenge 2019 dataset. We show significant performance boosts (absolute NDCG) by applying them in 4 representative baseline models: LF [9] (↑16.42%), HCIAE [23] (↑15.01%), CoAtt [39] (↑15.41%), and RvA [27] (↑16.14%). Impressively, on the official test-std server, we use an ensemble model of the most simple baseline LF [9] to beat the 2019 champion 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 performance.

2. Related Work

Visual Dialog. Visual Dialog [9, 11] is more interactive and challenging than most of the vision-language task, *e.g.*, image captioning [43, 42, 3] and VQA [4, 37, 36]. Specifically, Das *et al.* [9] collected a large-scale free-form visual dialog dataset VisDial [6]. 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 [11] 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 [16, 13, 34, 12, 33, 45]. They can be categorized by the usage of history. 1) Holistic: they treat history as a whole to feed into models like HACAN [41], DAN [18] and CorefNMN [20]. 2) Hierarchical: they use a hierarchical structure to deal with history like HRE [9]. 3) Recursive: RvA [27] 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 [23]: the first model to use history hierarchical attention, CoAtt [39]: the first one to a co-attention mechanism, and RvA [27]: the first one for a tree-structured attention mechanism.

Causal Inference. Recently, some works [17, 25, 10, 5] introduced causal inference into machine learning, trying to endow models the abilities of causal reasoning through the learning process. In contrast to them, we use the structural graph causality [29], 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 [29, 30].

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 $A_t = \{A_t^{(1)}, \ldots, A_t^{(100)}\}$, the task of the dialog agent is to generate a free-form answer or give an answer by ranking candidate answers A_t.

Evaluation. Recently, the ranking metric Normalized Discounted Cumulative Gain (NDCG) is adopted by the VisDial community. 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. More details of NDCG can be found in [1].

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 [29], 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 [32] based on ResNet backbone [14] pre-trained on Visual Genome [21]. For language feature, the encoder firstly embeds sentence into word vectors, followed by passing the RNN [15, 8] to generate features of question and history, which are denoted as $\{Q, H\}$.

Most of existing methods apply attention mechanism [40] 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 [23, 39, 41]. In the baseline graph, the sub-graph $\{I \rightarrow V, Q \rightarrow V, H \rightarrow Q \rightarrow V\}$ denotes a series of attention operations for visual knowledge V. Note that these arrows are not necessarily independent, such as co-attention [39], and the process can be written as $\textit{Input} : \{I, Q, H\} \Rightarrow \textit{Output} : \{V\}$, where intermediate variables can be yielded in the graph with respect to different attention strategies such as co-attention [39] and recursive attention [27]. However, without loss of generality, these variables 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 sentence that decoder takes the features via $\{H \rightarrow A, Q \rightarrow A, V \rightarrow A\}$ and then transforms them into a vector for decoding the answer. In particular, the decoder can be generative, *i.e*., to generate an answer sentence by RNN; or discriminative, *i.e*., select an answer by discriminating answer candidates.

Next, we will advance to the middle part of Figure 1, to reveal what is wrong with the baseline graph.

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 $R \rightarrow P$ because top researchers usually lead high-quality research, and $P \rightarrow A$ is even more obvious. Therefore, for the good of the community, the double-blind prohibits the direct link $R \rightarrow A$ 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 $H \rightarrow Q \rightarrow A$. If we inspect the role of H, we can find that it is to help Q resolve some co-reference 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: $H \rightarrow A$, 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 $H \rightarrow A$. 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 (*i.e.*, candidates with non-zero relevance score) average answer length represented as green dash line, and the other qualitative studies in Section 6.4.

4.2. Principle 2

Before discussing P2, we first introduce an important concept in causal inference [29]. 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.*, $Q \leftarrow U \rightarrow A$). The confounder U opens a non-causal path started from Q which is also called the *backdoor*, 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 the questioner and answerer can see the dialog history which offers them a latent topic, but also the answer annotators can look at the history when annotating the answer. Their preference can be understood as 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.

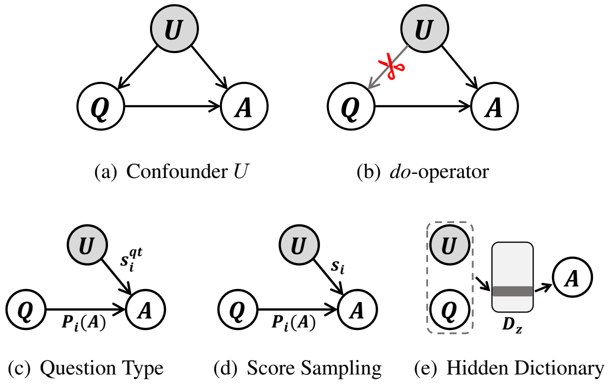


Figure 3. Example of confounder, *do*-operator and sketch causal graphs of our three attempts to de-confounder

It is worth noting that the confounder hinders us to find the true causal effect. Let's take the graph in Figure 3(a) as an example, if there is no U, the probability $P(A|Q)$ is the causal effect that we want to pursue. However, due to the existence of U, $P(A|Q)$ is no longer the true causality from Q to A. When we calculate $P(A|Q)$, we take U into account which can be shown by Bayes rule:

$$P(A|Q) = \sum\nolimits_u P(A|Q,u)P(u|Q). \tag{1}$$

The distribution of u is conditional on Q (*i.e.*, $P(u|Q)$). That means when using the conditional weight (*i.e.*, $P(u|Q)$) to sum every effect (*i.e.*, $P(A|Q,u)$), the likelihood sum (*i.e.*, $P(A|Q)$) will be biased towards the effect $P(A|Q,u)$ with larger weights. For better understanding, if we treat Eq (1) as a process of data stratification, at each layer u, we can obtain the causality conditional on u, because given u will block the backdoor of Q. Then, we have to sum these causalities by the natural distribution of u rather than conditional distribution $P(u|Q)$, which will remix the data bias. In a nutshell, we cannot calculate causality from Q to A by $P(A|Q)$ under the confounder U. To resolve this problem (*i.e.*, de-confounding to find causal effect), we need more powerful tools.

4.3. Overall Causal Graph

Here, we first introduce two additional tools: *do*-operator and *do*-calculus [29, 30], which can help us to de-confounder.

***do*-operator.** *do*-operator is a type of intervention to de-confounder. Illustrated in Figure 3(c), *do*-operator (*e.g.*, $do(Q = q)$) is that we set a value q to variable Q, *i.e.*, Q is caused by itself rather than its parent nodes. Therefore, *do*$(Q = q)$ cut off all the original arrows that come into Q (*i.e.*, $U \to Q$) because its parents do not cause it anymore. This operation can prevent any information about Q from flowing in the non-causal direction (*i.e.*, backdoor $Q \leftarrow U \to A$). As a result, the confounder of Q can be relieved and the causal effect of Q can be estimated. In the following parts, we use *do*(q) to represent *do*$(Q = q)$ for concision.

***do*-calculus.** However, it is hard to take a real intervention on a fixed dataset. We need to use some rules to translate $P(A|do(q))$ into $P(A|(Q,\dots))$, which has no *do*-operator and can be calculated by conditional probability. The rules of *do*-calculus are given in [29, 30] and here we just introduce the most important one: If a set Z of variables blocks all backdoor paths from X to Y, then conditional on Z, *do*(x) is equivalent to *observe*(x): $P(Y|do(x), Z) = P(Y|X, Z)$ where capital letter denotes variable and lowercase denotes value. Other rules will be given in supplementary materials.

After obtaining the tools, we can revisit the example in Section 4.2. If we calculate $P(A|do(q))$ rather than $P(A|Q)$, the result will be $\sum_u P(A|Q,u)P(u)$. In this formula, the distribution of u is the natural prior $P(u)$ instead of the conditional distribution $P(u|Q)$. Therefore, the summation of the causal effect by weight (*i.e.*, $P(u)$) will not remix the data bias. In other words, $P(A|do(q))$ is the ideal causality from Q to A.

In our graph of VisDial shown in Figure 1, we can also de-confounder U by intervention *do*(q, h, i) to find causal effects from $\{Q, H, I\}$ to A, then perform *do*-calculus rules to transform pretended intervention into probability formula:

$$\begin{aligned} &P(A|do(q,h,i)) \\ &= \sum\nolimits_u P(A|do(q,h,i),u)P(u|do(q,h,i)) \\ &= \sum\nolimits_u P(A|do(q),H,I,u)P(u|H) \\ &= \sum\nolimits_u P(A|Q,H,I,u)P(u|H). \end{aligned} \tag{2}$$

The last transformation takes the rule we introduced in *do*-calculus because Q's backdoors are blocked by controlling U. The rest derivation proofs and the details of other rules can be found in supplementary materials. As we mentioned, the result of $P(A|do(q,h,i))$ is the real causal effect that we want.

So far, we have given all of the contents about baseline causal graph, two principles and our causal graph. In the next section, we will try to calculate the real causal effect and give some attempts to realize our causal graph to enlight the future of visual dialog.

5. Improved Visual Dialog Models

It is easy to implement P1 and we will give some examples as training details in Section 6.3. As for P2, we can obtain causal effect estimation by Eq (2) which can be written as:

$$P(A|do(q,h,i)) = \sum\nolimits_u P_u(A)P(u|H), \tag{3}$$

where $P_u(A)$ represents the probability of A under the conditions Q, H, I and u. Since the variable U is unobserved, we just give some examples of attempts to replace U or approximate it and corresponding sketch graphs will be given to help understand.

5.1. Question Type

Inspired by data stratification form in Eq (3), we try to use question type to stratify the data. Specifically, we manually define some question types, count appeared answers and set preference for every answer in each type of question. According to the Eq (3), we can use the preference generated by question type to train our model with the loss function:

$$\mathcal{L}_{qt} = \sum\nolimits_i P_i(A) \cdot s_i^{qt}, \tag{4}$$

where i is the i-th candidate in answer list, $P_i(A)$ is the probability of candidate i, s_i^{qt} is the preference we counted and the sketch graph is shown in Figure 3(c). The implementation details will be given in Section 6.3.

5.2. Answer Score Sampling

The official gives a set of dense annotations in training set which can be treated as a representation of preference because the annotators score every candidate in the context H with their preference. As a result, if we regard each candidate A_i in the decoder as a u, illustrated in Figure 3(d), we can follow Eq (3) to calculate loss by the following function:

$$\mathcal{L} = -\sum\nolimits_i P_i(A) \cdot s_i, \tag{5}$$

where i is the index of answer candidate. Eq (5) can be implemented as different forms. Here we give three examples (detailed formulas are in supplementary materials):

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 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 is also corresponding relevance score.

Generalized Ranking Loss (R_3). Note that answer generation process can be viewed as a ranking problem. Therefore, we derive a ranking loss that $P_i(A)$ is a ranking probability $\log \frac{\exp(p_i)}{\exp(p_i)+\sum_{j\in G}\exp(p_j)}$ where G is a group of candidates which has a lower relevance score than i and s_i represents 0 (with no relevance score) or 1 (with positive relevance score). This loss function is reorganized from ListNet [7] to become more suitable for this task.

Note that our loss functions are derived from the Eq (3), not just the regression of dense annotation. The comparison experiments will be given in Section 6.4.

5.3. Hidden Dictionary Learning

We find that the Eq (3) can be written as:

$$\sum\nolimits_u P_u(A)P(u|H) = \mathbb{E}_{[u|H]}\left[P_u(A)\right]. \tag{6}$$

Although, we cannot determine the exact meaning of U, we try to use a vector representation $\boldsymbol{z}$ to approximate an expression of U. We can approximate $\mathbb{E}_{[U|H]}\left[P_u(A)\right]$ as $\text{NWGM}\left[P_u(A)\right]$ [40, 35] (*i.e.*, normalized weighted geometric mean), and this term can be further calculated by creating a dictionary D_z of $\boldsymbol{z}$:

$$\mathbb{E}_{[u|H]}\left[P_u(A)\right] \approx \text{Softmax}\{\boldsymbol{g}_z(\mathbb{E}_z\left[\boldsymbol{Z}\right])\}, \tag{7}$$

where $\boldsymbol{g}_z$ is a fully connected layer, $\boldsymbol{Z}$ represents a variable and its value $\boldsymbol{z}$ is selected from directory D_z. The details and proofs of the series of approximations can be found in supplementary materials. After deriving the last term, we can use D_z to calculate $\mathbb{E}_Z\left[\boldsymbol{Z}\right]$ shown in Figure 3(e) to approximate Eq (3). Noting that although when we train the dictionary, we still need to use answer score sampling, the hidden dictionary learning is a more proper way to approximate the unobserved confounder because it explores the whole space of U rather than the second attempt which only uses some samples of U.

6. Experiments

6.1. Experimental Setup

Dataset. Our principles are evaluated on the recently released real-world dataset VisDial v1.0[2]. Specifically, the training set of VisDial v1.0 contains 123K images from COCO dataset [22] with 10 rounds of dialog for each image, totally about 1.2M dialog pairs. 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 image in training and validation sets has a 10-round dialog, while in test set the number of the dialog is flexible. Every dialog in VisDial dataset is given with 100 answer candidates. We evaluated our results on the validation and test-std set.

Metric. We used Normalized Discounted Cumulative Gain (NDCG) to evaluate our models. As introduced in Section 3.1, NDCG is adopted as the new metric for visual dialog which is appointed by the official and accepted by the community. Note that 2018 and 2019 Visual Dialog challenge winners were both picked by NDCG.

6.2. Model Zoo

We report the performance of the following baseline VisDial models, including LF [9], HCIAE [23], CoAtt [39] and RvA [27]:

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

HCIAE [23]. The model consists of question-based history attention and question-history-based visual attention.

CoAtt [39]. 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 [27]. The model consists of question-based visual attention and history-based visual attention refinement.

6.3. Implementation Details

Pre-processing. As for language pre-processing, we followed the process introduced by [9]. Firstly, we lowercased all the letters in sentences, 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. And we built a vocabulary of the tokens of 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 [31] to initialize all word embeddings, which were shared in encoder and decoder, and we applied 2-layers LSTMs to encode word embedding and set its hidden states dimension to 512. As for the visual feature, we used bottom-up-attention features [3] given by the official [1].

Implementation of Principles. For P1, we eliminated the history feature in the final fused vector representation for all models, while kept other parts unchanged. For HCIAE [23] and CoAtt [39], we also blocked the history guidance to the image. For P2, we trained our models using the preference score, which was counted from question type or given by the official (*i.e.*, dense annotation in train v1.0). 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. "Answer score sampling" was directly used to train our pre-

[2] Suggest by the official [1], results should be reported on v1.0 instead of v0.9

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.

trained model by the proposed loss function. For "dictionary", we set a memory with the dimension 100×512 to realize D_z, then trained it by dense annotations with R_3 loss. More details can be found in supplementary materials. Note that other implementations following P1 and P2 are also acceptable.

Training. We used softmax cross-entropy loss to train the model with P1, and used Adam [19] with the learning rate of 4×10^{-3} which decayed at epoch 5, 7, 9 with the decay rate of 0.4. We trained the model for 15 epochs totally. Dropout [35] was also 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 attempts of P2 can further boost performance by 11.75% at most by hidden dictionary learning. To be more specific, our designed loss functions based on Eq. (3) outperform the regressive score (*i.e.*, R_0) which is a Euclidean distance loss, and we also find that our proposed generalized ranking loss (*i.e.*, R_3) is the best because it satisfies the ranking property of VisDial.

To justify that our principles are model-agnostic, Table 2 shows the results of our experiments about applying our principles on four different models (*i.e.*, LF [9], HCIAE [23], CoAtt [39] and RvA [27]). In general, both of our principles can improve all the models in any ablative conditions. We also find that the effectiveness of P1 and P2 are additive, that is to say, their combination performs the best. Note that the enhanced LF model is very simple without complex attention strategies. However, this simple architecture still does not hinder it to achieve the best performance.

6.5. Qualitative Analysis

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

Model	LF [9]	HCIAE [23]	CoAtt [39]	RvA [27]
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 attempts in Section 5 with the history shortcut.

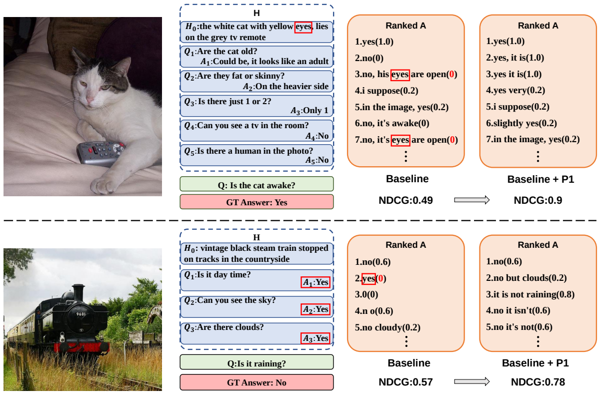


Figure 4. Qualitative results of the baseline and baseline with P1 on the validation set of VisDial v1.0. The numbers in brackets in ranked A denote relevance scores. Red boxes denote that the baseline model copies the words from the dialog history, even they are literally nonsense for answering the current question. The bottom example shows that although baseline can correctly select the ground truth answer, it is influenced by the unreasonable history shortcut to answer, and thus it ranks "yes" at a high place, which degrades its performance (NDCG). As for the baseline with P1, it does not make such unreasonable choices.

History Bias Elimination. After applying P1, many harmful patterns learned from history are relieved, especially the answer-length bias shown in Figure 2(a) and word-match bias shown in Figure 4. After applying P1, the average length of top-1 answers (*i.e.*, the blue line in Figure 2(a)) is no longer related to the history answer average length, and become more close to NDCG ground truth answer average length (*i.e.*, green dash line). As for the word-match bias in Figure 4, we can observe that the word "eyes" from history is literally unrelated to the current question. But in the top of the ranked answer list of the baseline model, the word "eyes" can be found in some undesirable candidates (*i.e.*, with low relevance score). In general, due to the wrong direct path from history to answer, the baseline model prefers to match the word in history and ranks matched candidates in high places. If we count the matching times of meaningful words on the validation set (*e.g.*, word "eyes") in the top-10 candidates of the ranked lists, obtained by baseline with P1 and the baseline, we find that P1 can decrease about

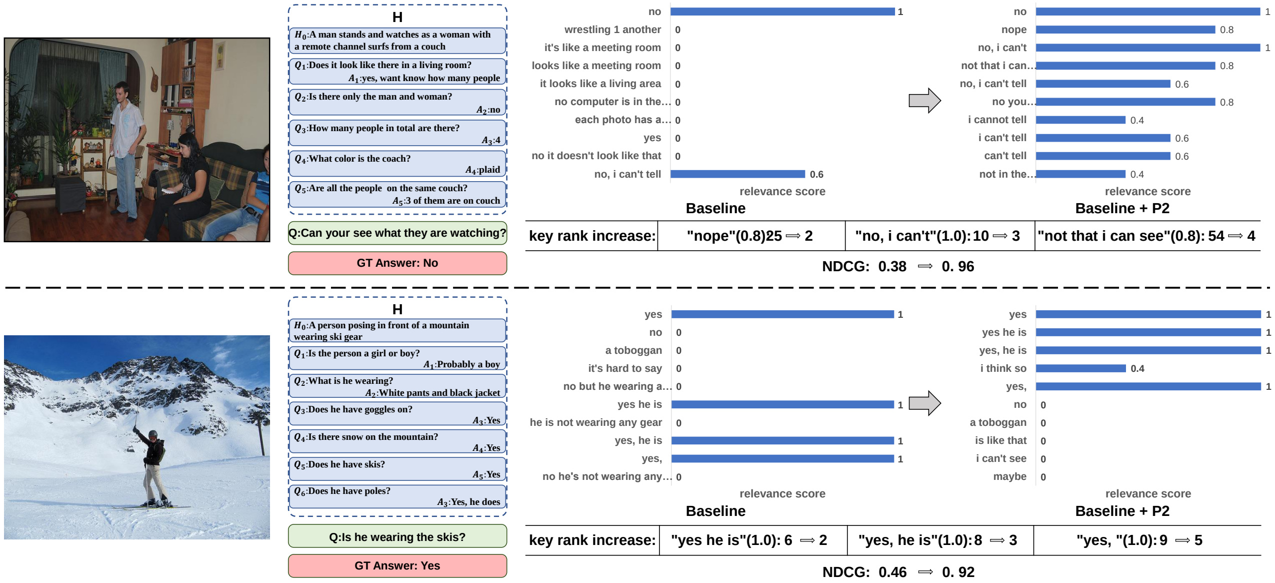


Figure 5. Qualitative examples of the ranked candidates of baseline and baseline with P2. We also give some key rank changes for boosting NDCG performance by implementing P2. These examples are taken from the validation set of VisDial v1.0.

	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	MReaL-BDAI*	74.02
	ReDAN+ (Ensemble) [12]	64.47
	square*	60.16
	VIC-SNU [18]*	57.59
	UET-VNU*	57.40
	idansc [33]*	57.13

Table 3. Our results and comparisons to the recent 2019 2nd Visual Dialog Challenge Leaderboard results on the test-std set of VisDial v1.0. Results are reported by the test server, (*) is taken from [2].

10% word matching from history ($\sim$ 4800 times compared with $\sim$ 5200 times).

The bottom example shown in Figure 4 also illustrates a type of word matching. In the ranked list of the baseline model, the rank of “yes” is very high, and “yes” exists in history for many times. By analyzing the results on validation, we found that if “yes” or “no” exists in dialog history, the baseline model will give the two answers a higher rank than average because of the word matching. After applying P1, this phenomenon will no longer happen. More details of these biases can be found in supplementary materials.

More Reasonable Ranking. Figure 5 shows that the baseline model only focuses on ground truth answer like “no” or “yes” and does not care about the rank of other answers with similar semantics like “nope” or “yes, he is”. This does not conform to human’s intuition because we think the candidates with similar semantics are still correct answers. This also leads the baseline model to perform badly under the NDCG metric. Compared with the model with P2, in the bottom example, it almost rank 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 greatly improves the NDCG performance.

6.6. Visual Dialog Challenge 2019

We finally used the blind online test server to justify the effectiveness of our principles on the test-std split of VisDial v1.0. Shown in Table 3, the top part contains the results of the baseline models implemented our principles, where P2 denotes the most effective one (*i.e.*, hidden dictionary learning). The bottom part is the 2019 Visual Dialog Challenge leader-board [2]. We used the ensemble of the enhanced LF [9] to beat our best performance in 2019 Visual Dialog Challenge, which also used other implementations of P1 and P2. Promisingly, by applying our principles, we can promote all the baseline single models to the top ranks on the leader-board.

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 implementation examples 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.

References

[1] Visual Dialog. https://visualdialog.org/. 1, 3, 6

[2] Visual Dialog Challenge 2019 Leaderboard. https://evalai.cloudcv.org/web/challenges/challenge-page/161/leaderboard/483/. 1, 8

[3] 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. 2, 6

[4] 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. 1, 2

[5] 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. *arXiv preprint arXiv:1901.10912*, 2019. 3

[6] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. Amazons mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5, 2011. 2

[7] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136. ACM, 2007. 5

[8] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014. 3

[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. 1, 2, 3, 6, 7, 8

[10] Ishita Dasgupta, Jane Wang, Silvia Chiappa, Jovana Mitrovic, Pedro Ortega, David Raposo, Edward Hughes, Peter Battaglia, Matthew Botvinick, and Zeb Kurth-Nelson. Causal reasoning from meta-reinforcement learning. *arXiv preprint arXiv:1901.08162*, 2019. 3

[11] 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. 2

[12] Zhe Gan, Yu Cheng, Ahmed EI Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. Multi-step reasoning via recurrent dual attention for visual dialog. *arXiv preprint arXiv:1902.00579*, 2019. 1, 2, 3, 8

[13] 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. 3

[14] 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. 3

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[16] 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. 2, 3

[17] Diviyan Kalainathan, Olivier Goudet, Isabelle Guyon, David Lopez-Paz, and Michèle Sebag. Sam: Structural agnostic model, causal discovery and penalized adversarial learning. *arXiv preprint arXiv:1803.04929*, 2018. 3

[18] Gi-Cheon Kang, Jaeseo Lim, and Byoung-Tak Zhang. Dual attention networks for visual reference resolution in visual dialog. *arXiv preprint arXiv:1902.09368*, 2019. 2, 3, 8

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[20] 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. 3

[21] 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. 3

[22] 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. 6

[23] 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. 1, 2, 3, 6, 7

[24] 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. 1

[25] Suraj Nair, Yuke Zhu, Silvio Savarese, and Li Fei-Fei. Causal induction from visual observations for goal directed tasks. *arXiv preprint arXiv:1910.01751*, 2019. 3

[26] 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. 1

[27] 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. 1, 2, 3, 6, 7

[28] Judea Pearl et al. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009. 1

[29] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. 1, 2, 3, 4, 5

[30] Judea Pearl and Dana Mackenzie. *THE BOOK OF WHY: THE NEW SCIENCE OF CAUSE AND EFFECT*. Basic Books, 2018. 3, 4, 5

[31] 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. 6

[32] 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. 3

[33] 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. 2, 3, 8

[34] 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. 3

[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. 6, 7

[36] 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. 2

[37] 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. 1, 2

[38] 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. 1

[39] 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. 1, 2, 3, 6, 7

[40] 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. 3, 6

[41] Tianhao Yang, Zheng-Jun Zha, and Hanwang Zhang. Making history matter: Gold-critic sequence training for visual dialog. *arXiv preprint arXiv:1902.09326*, 2019. 2, 3

[42] 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. 2

[43] 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. 2

[44] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multimodal 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. 1

[45] 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. 3

Supplementary Material

This supplementary material will give further details for the main paper, including A. More basic knowledge of causal graph for a better understanding of our graph, B. What the confounder brings to us and why we use *do*, C. Proofs and details which are omitted in the main paper due to space limitation, D. More qualitative examples to testify the effectiveness of our principles, E. The whole tables of our experiments.

A. Basic Knowledge of Causal Graph

A.1. Causal Graph

The basic definition of causal graph is introduced in the main paper. Here, it is beneficial for introducing more details of causal graph. The most naive configuration is $X \rightarrow Y$, which denotes X causes Y, or Y listens to X. This directed path from X to Y is called causal path, which denotes X's causal effect on Y. In the real world, what we want to know is the causal effect among variables, not just co-occurrence.

For easy to understand the theories we will introduce latter, we start from the simple causal graph configurations. There are three basic configurations in causal graph. 1) **Chain**—one arrow directed into and one arrow directed out of the middle variable—is shown in Figure 1(a). 2) **Fork**—two arrows emanating from the middle variable—is shown in Figure 1(b). 3) **Collider**—the middle variable receiving arrows from two other nodes—is like the configuration $X \rightarrow Z \leftarrow Y$, which is not shown in the picture because we will not use it.

A.2. Conditional independence

We introduce the dependency between variables in causal graph in this section. Using **Chain** shown in Figure 1(a) as an example, it is obvious that:

X and Z are dependent

i.e., for some x, z, $P(Z = z|X = x) \neq P(Z = z)$,

Z and Y are dependent

i.e., for some z, y, $P(Y = y|Z = z) \neq P(Y = y)$,

These two points are valid because according to the definition of causal graph, child node (*i.e.*, Z or Y) listens to its parent node (*i.e.*, X or Z) and decides its value in response to what it hears.

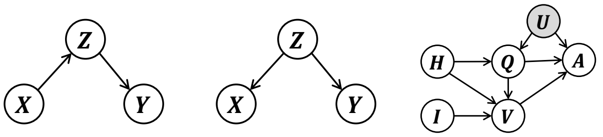
X and Y are likely dependent

i.e., for some x, y, $P(Y = y|X = x) \neq P(Y = y)$,

X and Y are independent, conditional on Z

i.e., for all x, y, z, $P(Y = y|Z = z, X = x) = P(Y = y|Z = z)$.

Here, we're only comparing cases where the value of Z is constant. Since Z does not change, the values of X and Y do not change in accordance with it. Therefore, any additional changes in the values of X and Y must be independent of each other. For example, we use X, Z, Y to represent three events "there is fire", "there is smoke" and "smoke detector is on" respectively. If Z is always equal to 1 (*e.g.*, "there is smoke" is always true), we will find that X will not influence Y, because whether "fire" is on, the event "smoke detector is on" is always true. Therefore, X and Y are independent conditional on Z. In the **Fork** shown in Figure 1(b), the conditional independence relationship among X, Y and Z is also satisfied (*i.e.*, X and Y are independent, conditional on Z).



(a) Chain (b) Fork (c) Our Graph

Figure 1. Examples of some causal graph configurations and our graph of visual dialog

B. Causal Effect, Confounder and *do*

In this section, we will give a systematical analysis of the influence of confounder, why we need *do* and how to calculate it. First, we need to give explanation of causal effect. Note that, in the following parts, capital letter denotes variable and lowercase denotes value.

B.1. Causal Effect

In the naive causal graph $X \rightarrow Y$, with no other nodes and arrows. The effect of X on Y should be $P(Y|X) - P(Y)$, but for the prior $P(Y)$, it is a constant. For convenience, in this paper, we sometimes use $P(Y|X)$ to represent the effect of X on Y. Because there is only a causal path from X to Y, the effect can only pass through the causal path. So, $P(Y|X)$ is the causal effect that we want to pursue. However, in the real world, things are not easy like this.

B.2. Confounder

The definition of confounder is introduced in Section 4 of the main paper. Use the **Fork** as an example, by the definition, we know that in Figure 1(b), Z is the confounder for X and Y. And in this graph, we know that X do not have causal effect on Y because there is no causal path from X to Y (*i.e.*, the causal effect of X on Y is 0).

When we calculate the causal effect of X on Y, in this graph, we find that we cannot use $P(Y|X)$ to represent causal effect. When we calculate $P(Y|X)$, the result contains a backdoor between them (*i.e.*, $X \leftarrow Z \rightarrow Y$), from where they pass correlation information. So, the result of

$P(Y|X) - P(Y)$ cannot be zero. More seriously, we do not have a mathematical notation to represent the causal effect, let alone calculating it.

In short, confounder makes us cannot use $P(Y|X)$ to represent the causal effect, and we need to use new notations to represent it.

B.3. *do*

In the book [8], they introduce a new notation $P(Y|do(X = x))$, which can be used to represent the causal effect of X on Y. In this section, we will introduce why it can represent the causal effect and how to calculate it. Note that we will use $do(x)$ to represent $do(X = x)$ for concision in the following sections.

***do*-operator** As we mentioned in the main paper, *do* is a type of intervention, which means that we set a value to the variable instead of that its parent nodes cause it. For example, in Figure 1(b), $do(x)$ is that we set the value x to X ignoring its caused function (*i.e.*, arrow $X \leftarrow U$). That means when we *do* a variable, we cut off all the arrows ending to the variable. As a result, when we calculate $P(Y|do(x))$, no non-causal path will be calculated, which ensures our results are all about causal effect. We will give an example in Section B.4 to testify the statement.

Now, although we have a notation for causal effect, we cannot calculate it by existing methods. We need to transform the *do* formula into probability formula. That is *do*-calculus.

***do*-calculus** Three rules of *do*-calculus are given in [8] to help us finish the transformation of eliminating *do*-operator.

Rule 1. When we observe a variable X that is irrelevant to Y (possibly conditional on other variables Z, like the example “Chain” in Figure 1(a)), then the probability distribution of Y will not change:

$$P(Y|z, X) = P(Y|z). \tag{1}$$

Rule 2. If a set Z of variables blocks all back-door paths from X to Y, then conditional on Z, like the example “Fork” in Figure 1(b), $do(x)$ is equivalent to *see*(x):

$$P(Y|do(x), z) = P(Y|X, z). \tag{2}$$

Rule 3. We can remove $do(x)$ from $P(Y|do(x))$ in any case where there are no causal paths from X to Y:

$$P(Y|do(x)) = P(Y). \tag{3}$$

B.4. Revisit the Fork

Now, we have *do*-operator to represent causal effect and *do*-calculus to calculate it. Let us revisit the problem bring by confounder in Section B.2. In Figure 1(b), $P(Y|do(X))$ can be further written as:

$$\begin{aligned} &P(Y|do(x)) \\ &= \sum\nolimits_z P(Y|do(x), z)P(z|do(x)) \\ &= \sum\nolimits_z P(Y|z)P(z|do(x)) \\ &= \sum\nolimits_z P(Y|z)P(z) \\ &= P(Y) \end{aligned} \tag{4}$$

The first line uses Bayes rules, the second one and third one use **Rule 3**. As a result, $P(Y|do(x)) - P(Y)$ is equal to 0, which accords with our inference of the causal effect of X on Y in Section B.2. That also means we can use $P(Y|do(X))$ to calculate causal effect.

In conclusion, confounder makes us cannot use $P(Y|X)$ to calculate the causal effect, and we obtain a new mathematical notation $P(Y|do(X))$ to denote it. For calculating *do* formula, we need *do*-calculus to transform the *do* formula into probability formula, which can be further calculated by observational data. That is the whole story of confounder and *do*.

C. Proofs and Details

C.1. Proofs of Equation 2

For convenience, we draw our graph in Figure 1(c) and write down the equation again, and add one intermediate step for the formula derivation:

$$\begin{aligned} &P(A|do(q, h, i)) \\ &= \sum\nolimits_u P(A|do(q, h, i), u)P(u|do(q, h, i)) \\ &= \sum\nolimits_u P(A|do(q, h, i), u)P(u|do(h)) \\ &= \sum\nolimits_u P(A|do(q), H, I, u)P(u|H) \\ &= \sum\nolimits_u P(A|Q, H, I, u)P(u|H). \end{aligned} \tag{5}$$

According to the rules of *do*-calculus introduced in Section B.3, we can derive the following proofs: The first step is according to the Bayes rules. The second one is due to Q, I do not have a causal path to U and **Rule 3**. Then, the third step is because H, I do not have a backdoor to A and **Rule 2**. As for the last step, although Q has two backdoors to A (*i.e.*, $Q \leftarrow H \rightarrow U \rightarrow A$ and $Q \leftarrow U \rightarrow A$), according to **Rule 2**, when we control U, all of the backdoors are blocked. As a result, the last transformation is valid.

C.2. Details of Loss Functions

Following the Equation 5 given in Section 5, we give three loss functions:

Weighted Softmax Loss(R_1).

$$R_1 = \sum\nolimits_i \log(\text{Softmax}(p_i)) \cdot s_i, \tag{6}$$

where p_i is the logit of candidate A_i, and s_i is the corresponding relevance score.

Binary Sigmoid Loss(R_2).

$$R_2 = \sum\nolimits_i \left[\log(\sigma(p_i)) \cdot s_i + \log(\sigma(1-p_i)) \cdot (1-s_i)\right], \tag{7}$$

where σ is the Sigmoid function, p_i is the logit of candidate A_i, and s_i is the corresponding relevance score.

Generalized Ranking Loss(R_3).

$$R_3 = \sum\nolimits_i \log \frac{\exp(p_i)}{\exp(p_i) + \sum_{j \in G} \exp(p_j)} \cdot s_i, \tag{8}$$

where p_i is the logit of candidate A_i, G is a group of candidates that has a lower relevance score than A_i. s_i equal to 1 when the corresponding relevance score greater than 0 and s_i equal to 0 when the corresponding relevance score equal to 0. Note that this function is reorganized from ListNet [2].

C.3. Proofs of Formula 7

According to [11], we can use $\mathrm{NWGM}\left[P_u(A)\right]$ (*i.e.*, normalized weighted geometric mean) to approximate $\mathbb{E}_{[U|H]}\left[P_u(A)\right]$. If the probability of z_i (*i.e.*, a value of Z) is $P(z_i)$, and because we do a softmax over the whole answer candidates, $P_u(A) \propto \exp(\boldsymbol{g}_z(z_i))$. We use n_i to denote $\boldsymbol{g}_z(z_i)$. $\mathbb{E}_{[U|H]}\left[P_u(A)\right]$ can be written as:

$$\begin{aligned} & \mathbb{E}_{[U|H]}\left[P_u(A)\right] \\ & \approx \mathrm{NWGM}\left[P_u(A)\right] \\ & = \frac{\prod_i \exp(n_{k,i})^{P(z_i)}}{\sum_j \prod_i \exp(n_{j,i})^{P(z_i)}} \\ & = \frac{\exp(\mathbb{E}_z[n_{k,i}])}{\sum_j \exp(\mathbb{E}_z[n_{j,i}])}, \end{aligned} \tag{9}$$

where k is the index of ground truth answer and j is the index of all the candidates. If $\boldsymbol{g}_z(\cdot)$ is a fully connected layer, the equation can be further written as:

$$\mathbb{E}_{[U|H]}\left[P_u(A)\right] \approx \mathrm{Softmax}\{\boldsymbol{g}_z(\mathbb{E}_z\left[\boldsymbol{Z}\right])\} \tag{10}$$

C.4. Details of Principle Implementation

Details of Enhanced LF[3]. After obtaining the vision and language feature $\mathcal{H}, \mathcal{Q}, \mathcal{I}$, we did the further operations (We use the notation Att to denote attention operation introduced in main paper): 1) History feature refine: $\tilde{\boldsymbol{h}} = Att(\mathcal{H}, \boldsymbol{q_t})$, where last term of $\mathcal{Q}$ (*i.e.*, $\boldsymbol{q_t}$) is guidance. 2) Question and caption feature refine: $\tilde{\boldsymbol{q}} = Att(\mathcal{Q}, \boldsymbol{c_t})$, $\tilde{\boldsymbol{c}} = Att(\mathcal{C}, \boldsymbol{q_t})$. 3) Vision feature refine: $\tilde{\boldsymbol{v}} = Att(\mathcal{V}, \{\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{c}}\})$. 4) Second step of vision feature refine: $\tilde{\boldsymbol{v'}} = Att(\tilde{\boldsymbol{v}}, \boldsymbol{g}_v([\tilde{\boldsymbol{h}}; \tilde{\boldsymbol{q}}]))$, where $\boldsymbol{g}_v$ is a fully connected layer followed by a Softmax function to generate weights for refining visual attention. 5) Feature fusion: $\boldsymbol{e} = \boldsymbol{g}_f([\tilde{\boldsymbol{v'}}; \tilde{\boldsymbol{q}}])$, where $\boldsymbol{g}_f$ is a multi-head fully connected layer. More details can be found in Table 1.

Details of P2. As to question type of P2, we manually defined 55 types of questions and then counted the occurrence of ground truth answers under these question types. We set answer candidates with occurrence greater than 5 as preferred answers and annotated their score as 1 under the corresponding question type. At training time, for obtaining a more reasonable preference, we did an approximation here. We pre-trained model by original methods for 5 epochs, and selected top-20 candidates of its prediction, and gave these selected candidates the relevance score we counted from question type for every round of each dialog. Then we used the refined QT-relevance score to further train our model by R_2. For using dense annotation by our loss function, we pre-trained the model for 5 epochs, and then further trained the models by these answer score sampling with our loss functions. As for the dictionary, we set a 100×512 dimension user dictionary to save the latent representation of U. We pre-trained the dictionary by one-hot ground truth answer, and then trained the dictionary by R_3 loss with the dense annotation. Then we fused the prediction of the dictionary and the prediction of pre-trained models by $logit_i + w \cdot d_i$, where $logit_i$ is the prediction of original models, d_i is the prediction of the dictionary, and w is a manually set weight which we set as 0.1. Then we further train the whole model by R_3.

D. More Qualitative Examples

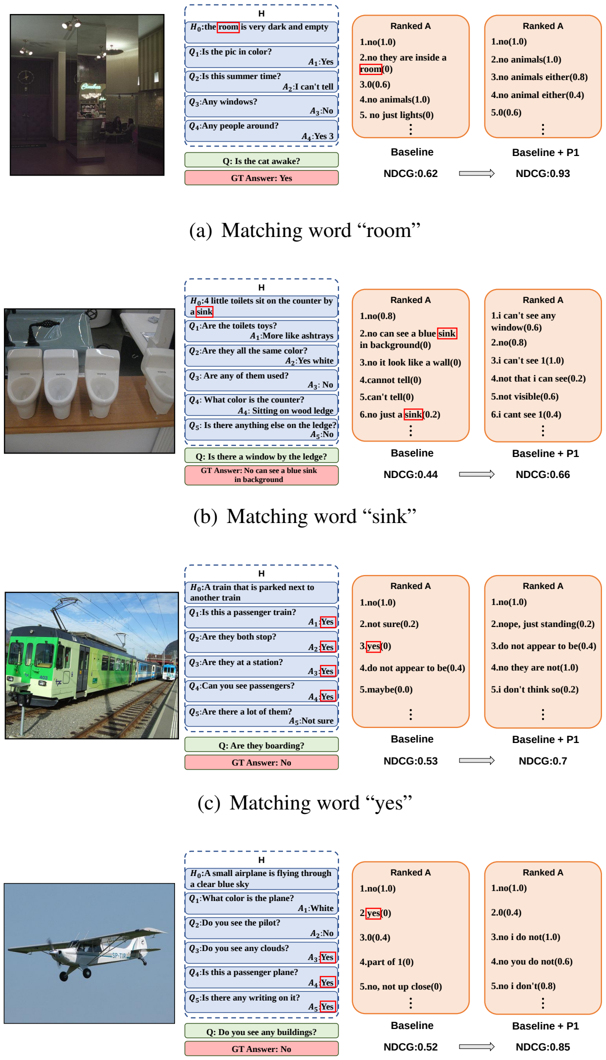
In this section, we will give more examples of the advantages of our principles mentioned in Section 6, including two types of history bias elimination for P1 shown in Figure 2(a) to Figure 2(d), and better ranking for P2 shown in Figure 3(a) and Figure 3(b).

E. The Whole Tables

In this section, we will give the whole tables of our experiments in the main paper, where we omit other metrics for concision. These metrics are: 1) mean rank of one-hot ground truth answer (*i.e.*, human response) (**Mean**), 2) recall@k (**R@k**), which is the existence of the human response in the ranked top-k candidates, 3) mean reciprocal rank (**MRR**) of the human response in the returned ranked list. Note that these old metrics are not suitable for visual dialog according to the suggestion of the official. Note that first and second VisDial challenge winners were both picked by **NDCG**.

References

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


H
H_0:the room is very dark and empty
Q_1:Is the pic in color?
A_1:Yes
Q_2:Is this summer time?
A_2:I can't tell
Q_3:Any windows?
A_3:No
Q_4:Any people around?
A_4:Yes 3
Q: Is the cat awake?
GT Answer: Yes
Ranked A
1.no(1.0)
2.no they are inside a room(0)
3.0(0.6)
4.no animals(1.0)
5. no just lights(0)
Ranked A
1.no(1.0)
2.no animals(1.0)
3.no animals either(0.8)
4.no animal either(0.4)
5.0(0.6)
Baseline
Baseline + P1
NDCG:0.62
NDCG:0.93

(a) Matching word “room”

H
H_0:4 little toilets sit on the counter by a sink
Q_1:Are the toilets toys?
A_1:More like ashtrays
Q_2:Are they all the same color?
A_2:Yes white
Q_3: Are any of them used?
A_3: No
Q_4: What color is the counter?
A_4: Sitting on wood ledge
Q_5: Is there anything else on the ledge?
A_5:No
Q: Is there a window by the ledge?
GT Answer: No can see a blue sink in background
Ranked A
1.no(0.8)
2.no can see a blue sink in background(0)
3.no it look like a wall(0)
4.cannot tell(0)
5.can't tell(0)
6.no just a sink(0.2)
Ranked A
1.i can't see any window(0.6)
2.no(0.8)
3.i can't see 1(1.0)
4.not that i can see(0.2)
5.not visible(0.6)
6.i cant see 1(0.4)
Baseline
Baseline + P1
NDCG:0.44
NDCG:0.66

(b) Matching word “sink”

H
H_0:A train that is parked next to another train
Q_1:Is this a passenger train?
A_1:Yes
Q_2:Are they both stop?
A_2:Yes
Q_3:Are they at a station?
A_3:Yes
Q_4:Can you see passengers?
A_4:Yes
Q_5:Are there a lot of them?
A_5:Not sure
Q: Are they boarding?
GT Answer: No
Ranked A
1.no(1.0)
2.not sure(0.2)
3.yes(0)
4.do not appear to be(0.4)
5.maybe(0.0)
Ranked A
1.no(1.0)
2.nope, just standing(0.2)
3.do not appear to be(0.4)
4.no they are not(1.0)
5.i don't think so(0.2)
Baseline
Baseline + P1
NDCG:0.53
NDCG:0.7

(c) Matching word “yes”

H
H_0:A small airplane is flying through a clear blue sky
Q_1:What color is the plane?
A_1:White
Q_2:Do you see the pilot?
A_2:No
Q_3:Do you see any clouds?
A_3:Yes
Q_4:Is this a passenger plane?
A_4:Yes
Q_5:Is there any writing on it?
A_5:Yes
Q: Do you see any buildings?
GT Answer: No
Ranked A
1.no(1.0)
2.yes(0)
3.0(0.4)
4.part of 1(0)
5.no, not up close(0)
Ranked A
1.no(1.0)
2.0(0.4)
3.no i do not(1.0)
4.no you do not(0.6)
5.no i don't(0.8)
Baseline
Baseline + P1
NDCG:0.52
NDCG:0.85

(d) Matching word “yes”

Figure 2. Word Matching

Index	Input	Operation	Output
(1)	H (word) (rnd ×40)	embed and LSTM	$\mathcal{H}$ (rnd ×512)
(2)	C (word) (1 × 20)	embed and LSTM	$\mathcal{C}$ (20 ×512)
(3)	Q (word) (1 × 20)	embed and LSTM	$\mathcal{Q}$ (20 ×512)
(4)	$(\mathcal{H}, q_t)$	Attention	$\tilde{h}$ (1 × 512)
(5)	$(\mathcal{C}, q_t)$	Attention	$\tilde{c}$ (1 × 512)
(6)	$(\mathcal{Q}, c_t)$	Attention	$\tilde{q}$ (1 × 512)
(7)	$(\mathcal{I}, \tilde{q}, \tilde{c})$	Attention	$\tilde{v}$ (2 × 2048)
(8)	$(\tilde{v}, [\tilde{h}; \tilde{q}])$	Attention	$\tilde{v}'$ (1 × 2048)
(9)	$(\tilde{v}, \tilde{q})$	Concatenate	$\boldsymbol{e}$ (1 × 2560)

Table 1. The details of Enhanced LF Encoder, where rnd is the current number round of history, C is image caption, $\mathcal{I}$ is the image feature offered by the official with the dimension 36 × 2048 and $\boldsymbol{e}$ is the output of encoder.

	P2	NDCG(%)	MRR(%)	R@1(%)	R@5(%)	R@10(%)	Mean
LF	baseline	57.12	64.33	50.46	81.41	90.15	4.03
	QT	58.97	64.42	50.70	81.40	89.93	4.13
	S(R_0)	67.82	51.82	40.66	63.31	75.86	8.21
	S(R_1)	71.27	51.40	38.30	65.54	78.78	7.09
	S(R_2)	72.04	50.84	38.65	63.54	77.76	7.26
	S(R_3)	72.36	50.38	37.13	64.22	78.09	7.13
	D	72.65	50.18	37.11	64.50	78.59	7.08
LF+P1	baseline	61.88	61.46	47.46	78.63	88.12	4.58
	QT	62.87	62.09	48.13	79.40	88.79	4.47
	S(R_0)	69.47	50.54	39.71	61.41	74.55	8.72
	S(R_1)	72.16	51.20	38.56	64.78	77.96	7.46
	S(R_2)	72.85	50.93	38.88	63.41	77.66	7.35
	S(R_3)	73.42	50.53	38.41	63.12	77.54	7.40
	D	**73.63**	50.56	37.99	63.98	77.95	7.26

Table 2. The whole table of 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.

Model	P	NDCG(%)	MRR(%)	R@1(%)	R@5(%)	R@10(%)	Mean
LF [3]	baseline	57.12	64.33	50.46	81.41	90.15	4.03
	+P1	61.88	61.46	47.46	78.63	88.12	4.58
	+P2	72.65	50.18	37.11	64.50	78.59	7.08
	+P1+P2	**73.63**	50.56	37.99	63.98	77.95	7.26
HCIAE [6]	baseline	56.98	64.13	50.31	81.42	90.18	4.09
	+P1	60.12	61.00	46.66	78.74	88.34	4.61
	+P2	71.50	46.96	32.43	63.47	78.43	7.28
	+P1+P2	71.99	46.83	33.20	61.64	76.53	7.67
CoAtt [10]	baseline	56.46	63.81	49.77	81.20	90.19	4.13
	+P1	60.27	60.97	46.83	78.29	87.86	4.66
	+P2	71.41	47.32	33.35	63.51	77.26	7.56
	+P1+P2	71.87	46.41	32.79	61.27	76.37	7.87
RvA [7]	baseline	56.74	64.49	50.67	81.64	90.50	3.98
	+P1	61.02	62.00	47.99	79.14	89.04	4.42
	+P2	71.44	50.33	36.85	64.94	78.81	7.05
	+P1+P2	72.88	49.34	36.62	62.96	77.75	7.44

Table 3. The whole table 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 2. Note that only applying P2 is implemented by the attempts in Section 5 in main paper with the history shortcut.

[2] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136. ACM, 2007.

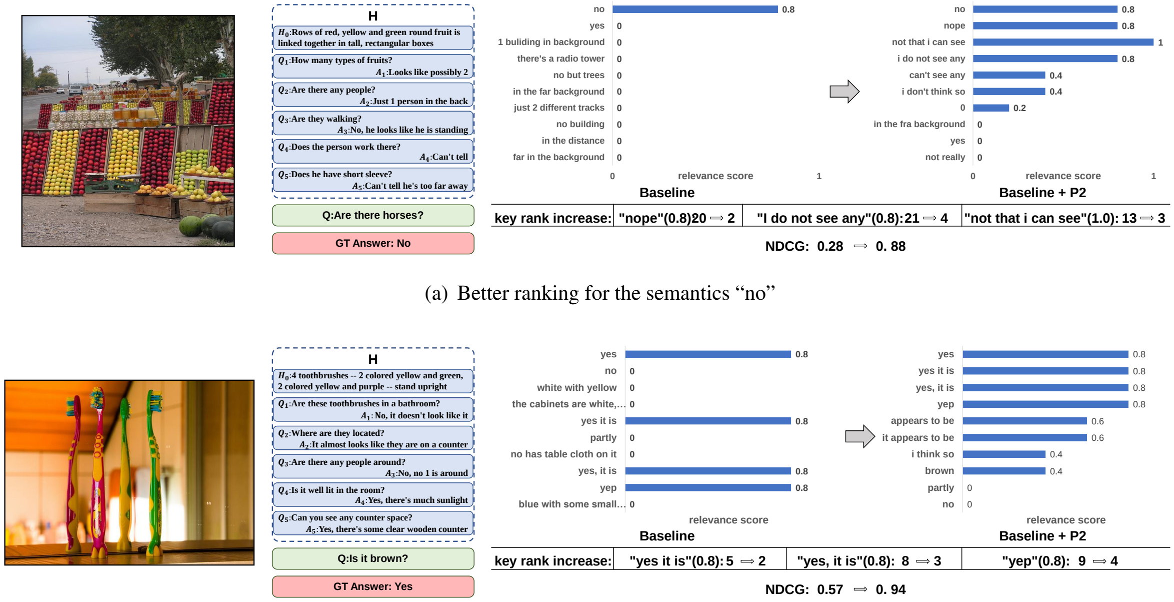
[3] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335, 2017.

[4] Zhe Gan, Yu Cheng, Ahmed EI Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. Multi-step reasoning via recurrent dual attention for visual dialog. *arXiv preprint arXiv:1902.00579*, 2019.

[5] Gi-Cheon Kang, Jaeseo Lim, and Byoung-Tak Zhang. Dual attention networks for visual reference resolution in visual dialog. *arXiv preprint arXiv:1902.09368*, 2019.

[6] 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.

[7] Yulei Niu, Hanwang Zhang, Manli Zhang, Jianhong Zhang, Zhiwu Lu, and Ji-Rong Wen. Recursive visual attention



(a) Better ranking for the semantics "no"

(b) Better ranking for the semantics "yes"

Figure 3. Better Ranking

	Model	NDCG(%)	MRR(%)	R@1(%)	R@5(%)	R@10(%)	Mean
Ours	P1+P2 (More Ensemble)	**74.91**	49.13	36.68	62.96	78.55	7.03
	LF+P1+P2 (Ensemble)	74.19	46.69	32.45	62.13	77.10	7.33
	LF+P1+P2 (single)	71.60	48.58	35.98	62.08	77.23	7.48
	RvA+P1+P2 (single)	71.28	47.71	34.80	61.53	77.10	7.63
	CoAtt+P1+P2 (single)	69.81	44.83	30.83	60.65	75.73	8.08
	HCIAE+P1+P2 (single)	69.66	44.03	29.85	59.50	75.98	8.10
Leaderboard	MReaL-BDAI*	74.02	52.62	40.03	68.85	79.15	6.76
	ReDAN+ (Ensemble) [4]	64.47	53.75	42.45	64.68	75.68	6.63
	square*	60.16	61.62	47.15	78.73	88.48	4.46
	VIC-SNU [5]*	57.59	63.22	49.60	79.73	89.15	4.31
	UET-VNU*	57.40	59.50	45.50	76.33	85.82	5.43
	idansc [9]*	57.13	69.25	55.65	86.73	94.05	3.14

Table 4. Our results and comparisons to the 2019 2nd Visual Dialog Challenge Leaderboard results on the test-std set of VisDial v1.0. Results are reported by the test server, (*) is taken from [1].

in visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6679–6688, 2019.

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

[9] 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.

[10] 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.

[11] 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.