Context-Aware Graph Inference with Knowledge Distillation for Visual Dialog

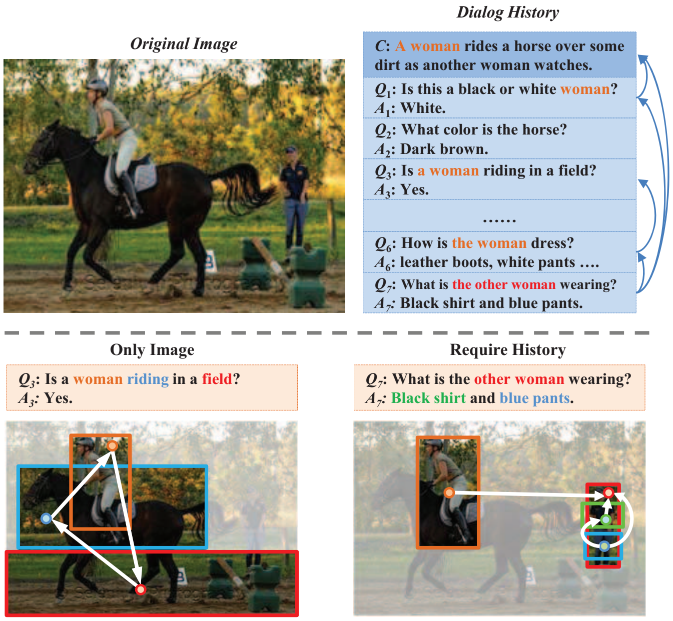
Dan Guo, Hui Wang, and Meng Wang, *Fellow, IEEE*

Abstract—Visual dialog is a challenging task that requires the comprehension of the semantic dependencies among implicit visual and textual contexts. This task can refer to the relational inference in a graphical model with sparse contextual subjects (nodes) and unknown graph structure (relation descriptor); how to model the underlying context-aware relational inference is critical. To this end, we propose a novel Context-Aware Graph (CAG) neural network. We focus on the exploitation of fine-grained relational reasoning with object-level dialog-historical co-reference nodes. The graph structure (relation in dialog) is iteratively updated using an adaptive top-K message passing mechanism. To eliminate sparse useless relations, each node has dynamic relations in the graph (different related K neighbor nodes), and only the most relevant nodes are attributive to the context-aware relational graph inference. In addition, to avoid negative performance caused by linguistic bias of history, we propose a pure visual-aware knowledge distillation mechanism named CAG-Distill, in which image-only visual clues are used to regularize the joint dialog-historical contextual awareness at the object-level. Experimental results on VisDial v0.9 and v1.0 datasets show that both CAG and CAG-Distill outperform comparative methods. Visualization results further validate the remarkable interpretability of our graph inference solution.

Index Terms—Visual dialog, cross-modal interaction, relational reasoning, graph inference, knowledge distillation

1 Introduction

Recently, cross-modal semantic reasoning between vision and language has attracted more and more interests, such as referring expression [1], [2], [3], [4], [5], visual captioning [6], [7], [8], [9], [10], [11], visual question answering (VQA) [12], [13], [14], [15], [16], [17], and image/video-text retrieval [18], [19] involving both images and videos. In these works, semantic referring between vision and language is always performed in a one-way single round. Taking VQA as an example, the agent identifies regions of interest in an image related to a specific question and infers an answer. In this work, we focus on a more challenging cross-modal interaction task, *i.e.*, visual dialog accompanied by multi-round question-answer (QA) pairs [20], [21], [22]. In the visual dialogue task, the interaction between the image and historical conversation is progressively changing, and the relationships among various objects in the image are influenced by the current question. During the dialogue, how to acquire question-conditioned context awareness has to be addressed. The exploitation of both visual and textual contexts benefits relational reasoning. The core technical point is to dig out underlying dynamic semantic dependencies in and between textual and visual contexts.



C: A woman rides a horse over some dirt as another woman watches.
Q_1: Is this a black or white woman?
A_1: White.
Q_2: What color is the horse?
A_2: Dark brown.
Q_3: Is a woman riding in a field?
A_3: Yes.
……
Q_6: How is the woman dress?
A_6: leather boots, white pants ….
Q_7: What is the other woman wearing?
A_7: Black shirt and blue pants.

Q_3: Is a woman riding in a field?
A_3: Yes.

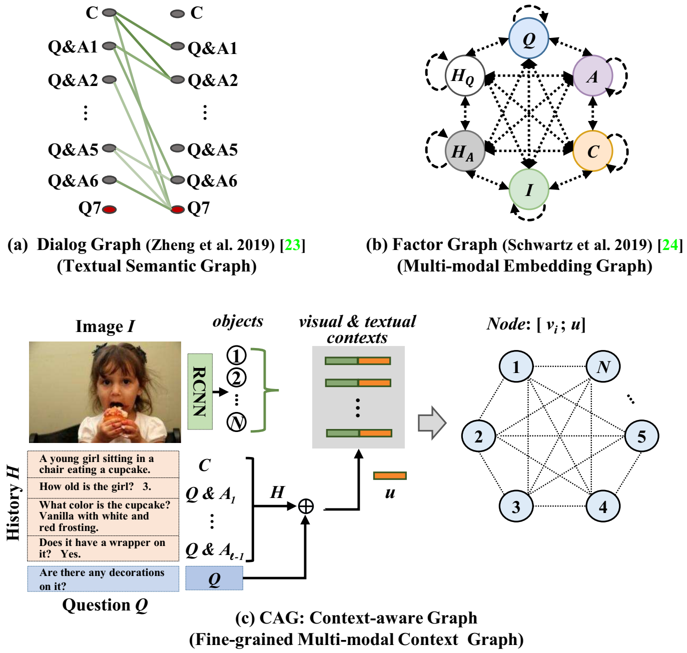
Q_7: What is the other woman wearing?
A_7: Black shirt and blue pants.

Fig. 1. Two challenges of visual dialog: (1) rich visual clues involve fine-grained relational inference and (2) when and how to explore historical clues? The resolution of the visual-textual contextual correlation - the proposed context-aware graph - CAG - are depicted in Figs. 2 (c) and 3.

For contextual awareness learning, as shown in Fig. 1, there are two to-be-solved issues as follows. **(1)** Fine-grained visual context (*i.e.*, relation among object-level entities) contributes informative clues for inferring answer. In order to better understanding the dynamic relation in the dialogue, the graph structure is employed to exploit the co-reference of image and history under the guidance of question. As shown in Fig. 2, prior graph-based models refer to different structures with textual nodes in [23] (question-answer pair at each round) and multi-modal nodes in [24] (embedding

- *D. Guo, H. Wang, and M. Wang are with Key Laboratory of Knowledge Engineering with Big Data (HFUT), Ministry of Education, School of Computer Science and Information Engineering School of Artificial Intelligence, Hefei University of Technology (HFUT), and Intelligent Interconnected Systems Laboratory of Anhui Province (HFUT), Hefei, 230601, China. E-mail: guodan@hfut.edu.cn, wanghui.hfut@gmail.com, and eric.mengwang@gmail.com.*
(D. Guo and H. Wang contribute equally to this work. Corresponding author: M. Wang.)
Our code is available at: https://github.com/wh0330/VisDial-CAG-Distill

Manuscript received XX XX, 2020; revised XX XX, 2020.


C
Q&A1
Q&A2
Q&A5
Q&A6
Q7
C
Q&A1
Q&A2
Q&A5
Q&A6
Q7
Q
H_Q
A
H_A
C
I
Image I
objects
RCNN
visual & textual contexts
Node: [v_i ; u]
1
2
3
4
5
N
History H
A young girl sitting in a chair eating a cupcake.
How old is the girl? 3.
What color is the cupcake? Vanilla with white and red frosting.
Does it have a wrapper on it? Yes.
Are there any decorations on it?
Question Q
C
Q & A_1
Q & A_{t-1}
H
Q
u

(a) Dialog Graph (Zheng et al. 2019) [23] (Textual Semantic Graph)

(b) Factor Graph (Schwartz et al. 2019) [24] (Multi-modal Embedding Graph)

(c) CAG: Context-aware Graph (Fine-grained Multi-modal Context Graph)

Fig. 2. Different graph structures of existing graph works and ours. [23] (a) focused on edge learning (textual inference) with nodes of caption and question-answer pair at each round. [24] devoted to the semantic interaction of the six multi-modal entities (image I, caption C, the current question Q, option answer A, questions in history H_Q and answers in history H_A). Our solution (c) focuses on a more fine-grained context-aware graph, devoting to learning object-level dialog-historical contextual awareness under the guidance of question.

vectors of image I, caption C, the current question Q, option answer A, questions in history H_Q and answers in history H_A). For relational learning, [23] solved multi-round textual inference with question Q and history H and [24] modeled the multi-modal interaction of feature embedding of Q, H, image I, and A. In contrast, in our work, we focus on the exploitation of fine-grained relational reasoning, *i.e.*, object-level dialog-historical co-reference nodes. **(2)** Whether textual context in history is requisite for every question in the conversation? Question Q_3 in Fig. 1 can be answered with only an image, whereas the answer of question Q_7 has to be inferred from both visual and historical clues. Moreover, there is another fact that various attention [25], [26], [27], fusion [21], [28], and gating [29] tactics fall into spurious probability learning - cannot eliminate the linguistic bias of history. As the Joint CAG model shown in Fig. 3, it introduces irrelevant historical noises. It prefers the candidate answers related to high-frequency words "No" in QA pairs, whereas "Yes, like pros" is the ground-truth answer. We attempt to refine the "spurious" visual-textual relational knowledge learned in the task.

In order to explore complicated semantic dependencies in and between textual and visual contexts, we propose a Context-Aware Graph (CAG) neural network.

Solution I - CAG. For the fine-grained relation modeling, as shown in Fig. 2 (c), each node in our graph is a joint context representation, which contains both visual-objects and textual-history contexts; each edge involves the fine-grained visual interaction among various scene objects in the image. In our graph solution, all the nodes and edges are iteratively updated through an adaptive top-K message passing mechanism. To be specific, to eliminate sparse useless relations, our CAG solution is an asymmetric dynamic directed-graph, which models adaptive message passing in this relational graph. As shown in Fig. 4, at each message passing turn, each node adaptively selects the K most relevant nodes and only receives messages from these neighbors. The iterative graph inference process is similar to imitating humans checking out the implicit clues multiple times. Finally, after multi-turns graph inference, we impose graph attention on all the nodes to learn the final graph embedding for the answer inferring.

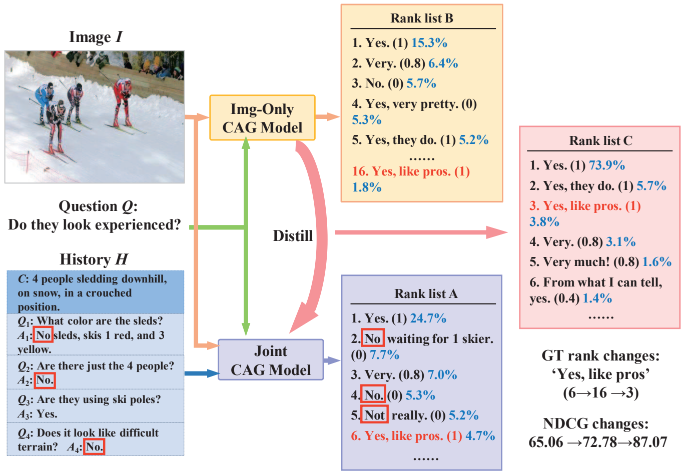


Image I
Question Q:
Do they look experienced?
History H
C: 4 people sledding downhill, on snow, in a crouched position.
Q_1: What color are the sleds?
A_1: No sleds, skis 1 red, and 3 yellow.
Q_2: Are there just the 4 people?
A_2: No.
Q_3: Are they using ski poles?
A_3: Yes.
Q_4: Does it look like difficult terrain? A_4: No.
Img-Only CAG Model
Joint CAG Model
Distill
Rank list B
1. Yes. (1) 15.3%
2. Very. (0.8) 6.4%
3. No. (0) 5.7%
4. Yes, very pretty. (0) 5.3%
5. Yes, they do. (1) 5.2%
......
16. Yes, like pros. (1) 1.8%
Rank list A
1. Yes. (1) 24.7%
2. No waiting for 1 skier. (0) 7.7%
3. Very. (0.8) 7.0%
4. No. (0) 5.3%
5. Not really. (0) 5.2%
6. Yes, like pros. (1) 4.7%
......
Rank list C
1. Yes. (1) 73.9%
2. Yes, they do. (1) 5.7%
3. Yes, like pros. (1) 3.8%
4. Very. (0.8) 3.1%
5. Very much! (0.8) 1.6%
6. From what I can tell, yes. (0.4) 1.4%
......
GT rank changes:
'Yes, like pros'
(6→16 →3)
NDCG changes:
65.06 →72.78→87.07

Fig. 3. Negative influences of historical clues in visual dialog. The format of answer in each rank list is fixed as "predicted rank order - answer - (annotated relevance score) - predicted probability". In answer list A, the joint model prefers candidate answers related to high-frequency words in QA pairs. To improve this linguistic bias, we apply the generalized visual knowledge in the image-only model (answer list B) to distill the joint model and output the final answer list C. The distilled knowledge is helpful to inhibit the historical noise, and makes the model much more self-confident, *e.g*, the rank 1 answer in list C with high predicted probability score.

Solution II - CAG-Distill. In the task, history is always deemed as an auxiliary to visual learning to identify the subject in the conversation, such as pronouns (*e.g.*, "it", "they", "he"); we have already added textual features onto each node in Solution I. Here, we discuss the non-necessity and linguistic bias of history. If the learned textual representation is needless or noisy, it thus leads to suboptimal performance. To address this issue, we introduce a two-branch optimization framework, where a visual-only (Img-only) CAG model captures pure object-level visual interaction as privileged information, and then injects it into a Joint dialog-historical CAG model by performing knowledge distillation [30], [31] between the answer logits. Compared with state-of-the-art approaches that impose hard constraints on answer decoding, our proposed method applies soft regularization on logits (*i.e.*, soft labels), which thus makes the relational graph learning more robust. To our knowledge, it is the first work to consider the knowledge distillation in the visual dialog task, which can inspire other cross-modal interaction tasks. We refer to this mechanism as visual-aware knowledge distillation. During testing, only the Joint CAG model is used, which leverages the distilled relation with only vision embedded. The essence is that a good model is not designed to fit hard labels from training data, but to learn the generalization of the data. After learning the generalized visual-aware ability from the teacher model (Img-only), the student model (Joint) benefits better results than without distillation, especially while the entropy of soft targets is higher than that of hard targets in our task.

The main contributions of the work are summarized as follows:

- We propose a Context-Aware Graph (CAG) neural network for visual dialog, which targets to discover partially relevant contexts and build the appropriate graph structure (dynamic relation). Actually, our graph CAG is an asymmetric dynamic directed-graph, and the solution is an adaptive inference process. The graph learning considers much more flexible, effective, and relevant message propagation.
- We propose a visual-aware knowledge distillation mechanism to address the issue of noisy historical context feature learning that exists in the task. We aim to obtain generalized awareness capacity from both visual and textual contexts.
- Extensive experiments are conducted on VisDial v0.9 and v1.0 datasets and achieve new state-of-the-art performances among graph-based methods for visual dialog.

The CAG approach was first introduced in our previous work [32]. Compared to the preliminary version, in this paper, we have made improvements in four aspects: (1) CAG merely solves the first challenge in Fig. 1; in this manuscript, we address both challenges by using knowledge distillation on CAG (CAG-Distill) as introduced in Sec. 1; (2) we perform a more comprehensive survey of existing related works, e.g., adding the review of knowledge distillation works in Sec. 2.3; (3) we add a new optimization - knowledge distillation on CAG - in Sec. 4; (4) we conduct more empirical evaluations and more discussions and analyses are provided in Sec. 5. In brief, although there are some literal overlaps, the new content in this manuscript makes the proposed graph framework CAG much more general, comprehensive, and convincing.

The remainder of this paper is organized as follows. Sec. 2 provides an overview of related works. Sec. 3 elaborates on the proposed graph model - CAG - for relational inference. Sec. 4 introduces a distillation scheme to eliminate the linguistic bias of history. The analysis of the experimental results is presented in Sec. 5, and conclusions are given in Sec. 6.

2 Related work

2.1 Visual Dialog

The visual dialog task has been introduced in recent years [20], [21], [22]. Plenty of methods based on the encoder-decoder framework have been introduced for visual dialog. Current encoder-based works can be divided into three facets. (1) **Fusion-based model.** Late fusion (LF) and hierarchical recurrent network (HRE) were introduced in [21]. These methods directly encoded the multi-modal inputs and fused them at different stages. (2) **Attention-based model**. To improve performance, attention mechanisms have been widely used in the task, including history-conditioned image attention (HCIAE) [33], sequential co-attention (CoAtt) [34], dual visual attention (D-VAN) [35], recurrent dual attention (ReDAN) [25], textual-visual reference-aware attention (RAA-Net) [26], and modular co-attention (MCA) [27]. (3) **Visual co-reference resolution model**. There are some attention models focus on explicit visual co-reference resolution. Seo et al. [36] designed an attention memory (AMEM) to store previous visual attentions. Kottur et al. [37] utilized neural module networks [38] to handle visual co-reference resolution at word-level. Niu et al. [39] proposed a recursive visual attention (RvA) mechanism to recursively reviews history to refine visual attention. Until now, existing methods have achieved a great process in the visual dialog community. However, besides discovering and attending important visual and textual clues, by characteristics of the task, relational reasoning and common sense reasoning in the conversion have become great interests in the field. More and more works turn to focus on reasoning aspects.

2.2 Graph Inference (GNN)

Graph neural networks have attracted attention in various vision tasks [40], [41], [42], [43], [44]. The core idea is to combine the graphical structural representation with neural networks, which is suitable for reasoning-style tasks. Liu et al. [45] proposed the first GNN-based approach for **VQA**, which applied external knowledge to build a scene graph and parse the questions. Later, Norcliffe-Brown et al. [46] modeled a graph representation conditioned on the question, and exploited a novel graph convolution to capture the interactions between different nodes. As for **visual dialog**, there are merely a few related works. Zheng et al. [23] proposed an EM-style GNN, it merely regarded the previous dialog-history as observed nodes, and the answer was deemed as unobserved node that can be inferred using EM algorithm on the textual context. Schwartz et al. [24] proposed a factor graph attention mechanism, which constructed the graph over all the multi-modal features and estimated their attention interactions. Jiang et al. [29] constructed a fully-connected scene graph over the image; they used a pre-trained visual relationship encoder [47] to directly learn the relations among the visual objects and employed a question-guided graph convolution for graph representation learning.

Fig. 2 illustrates the difference between graph models [23], [24] and our work. Apart from the graph structure referring to different multi-modal entities (nodes) and relational edges as shown in Fig. 2, technically, prior graph-based models considered the fixed graph attention or embedding, such as *fixed* fully-connected graph (FGA [24] and DualVD [29]), *fixed* once graph evolution (FGA [24] and DualVD [29]) and *fixed* unidirectional message passing (GNN [23]). In this paper, we are inspired by the nature of the visual dialog task, *i.e.*, multi-modal co-references in multi-round conversations. Fig. 4 shows the flexibility and adaptivity of our graph-based method, which iteratively evolves by *adaptive top-K* and *adaptive-directional* message passing. The significance of our method is that it exploits

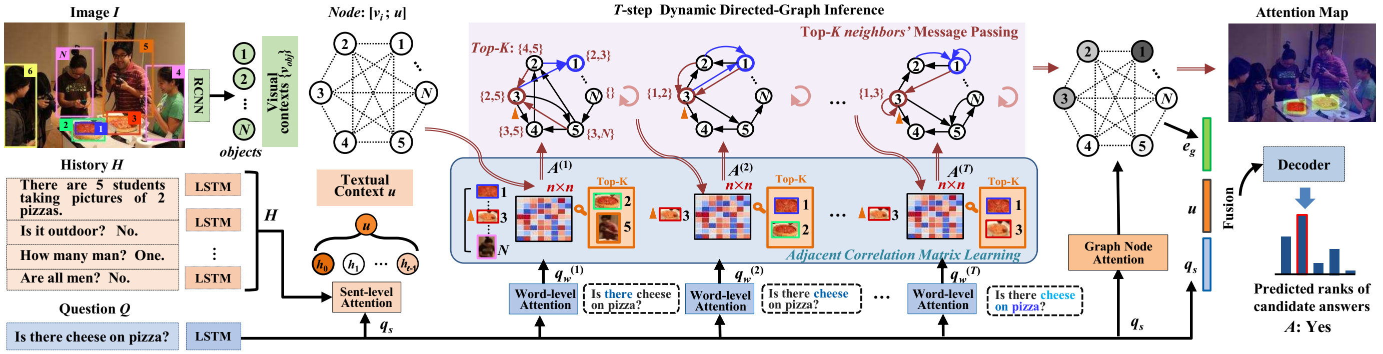


Image I
Node: $[v_i; u]$
T-step Dynamic Directed-Graph Inference
Attention Map
Top-K neighbors' Message Passing
Top-K: {4,5}
{2,3}
{2,5}
{3,5}
{3,N}
{}
{1,2}
{1,3}
RCNN
Visual contexts $\{v_{obj}\}$
objects
History H
There are 5 students taking pictures of 2 pizzas.
Is it outdoor? No.
How many man? One.
Are all men? No.
LSTM
H
Textual Context u
u
h_0 h_1 ... $h_{\ell-1}$
Sent-level Attention
q_s
Question Q
Is there cheese on pizza?
$A^{(1)}$ $A^{(2)}$ $A^{(T)}$
$n \times n$
Top-K
Adjacent Correlation Matrix Learning
$q_w^{(1)}$ $q_w^{(2)}$ $q_w^{(T)}$
Word-level Attention
Graph Node Attention
e_g
u
q_s
Fusion
Decoder
Predicted ranks of candidate answers
A: Yes

Fig. 4. The overall framework of Context-Aware Graph. Our context-aware graph is constructed with visual contexts $\{v_{obj}\}$ and textual context u. The dynamic relation between the nodes is iteratively inferred via Top-K neighbors' Message Passing under the guidance of word-level question command $q_w^{(t)}$. For example, the red and blue nodes in the graph respectively have different top-2 related neighbor nodes, and different directions of the message passing flow on the connection edges.

the image-history co-reference in a dynamic adaptive graph learning mode (*i.e.*, dynamic relation learning).

2.3 Knowledge Distillation

Knowledge distillation was firstly introduced for model compression [30], [48]. It proposed a training procedure to transfer knowledge from a pre-trained large model or ensemble of models to a small model, thus distilling knowledge from a heavier to a lighter model. Later, Lopez-Paz et al. [49] designed a knowledge distillation scheme to introduce privileged information [50], which some additional information that was only used in the training phase and excluded during testing. One application of this approach was [51], it treated the well-labeled modality as the privileged information and transferred the knowledge to the unlabeled modality for representation learning. [52] minimized the distillation loss between the new (hidden states) and the pre-trained activations.

In the visual dialogue task, we observe that the models using or without using dialog-history information have different appearances for answer prediction, while the visual awareness in these two cases should have common latent pattern. How to effectively utilize these two types of model setting to generalize the learning ability of model is a key challenge. In our solution, we innovatively regard the knowledge learned in the Img-only CAG model as the privileged information. Then, in the training phase, we distill this knowledge to the Joint CAG model to boost the model's generalization ability. The distilled Joint CAG model (CAG-Distill) is executed for answer prediction.

3 Method: Context-Aware Graph Inference

In this section, we introduce the proposed CAG (Context-Aware Graph) network. The visual dialog task refers to relational learning, which involves complicated underlying context dependencies between image, question and history. How to model the context-aware reasoning is critical. In this paper, we propose a dynamic graph inference to iteratively review the context clues. Given an image I and dialog history $H = \{C, (q_1, a_1), ..., (q_{\ell-1}, a_{\ell-1})\}$, where C is the image caption, (q, a) is a question-answer pair, and ℓ is the turn number of current dialog. The goal of the model is to infer an answer for the current question Q by ranking a list of 100 candidate answers $A = \{a_\ell^{(1)}, ..., a_\ell^{(100)}\}$. The following sub-sections describe the details of the CAG model.

Fig. 4 provides an overview of the proposed CAG. Specifically, CAG consists of three components: (1) **Graph Construction** (Sec. 3.1), which constructs the context-aware graph based on the representations of dialog-history and objects in the image; (2) **Iterative Dynamic Directed-Graph Inference** (Sec. 3.2), the context-aware graph is iteratively updated via T-steps dynamic directed-graph inference; (3) **Graph Attention Embedding** (Sec. 3.3), which applies a graph attention to aggregate rich node semantics. Then we jointly utilize the generated graph, the encoded question and history context features to infer the final answer.

3.1 Graph Construction

3.1.1 Feature Representation.

Given an image I, we extract the object-level features using Faster-RCNN [53] and apply a single-layer MLP with activation *tanh* to encode them into $V = \{v_1, ..., v_n\} \in \mathbb{R}^{d \times n}$, where n is the number of detected objects. For the current question Q, we first transform it into word embedding vectors $\mathcal{W}^Q = (w_1, ..., w_m) \in \mathbb{R}^{d_w \times m}$, where m denotes the number of tokens in Q. Then we use an LSTM to encode $\mathcal{W}^Q$ into an encoding sequence $U^Q = (h_1^q, ..., h_m^q) \in \mathbb{R}^{d \times m}$, and take the last vector h_m^q as the sentence-level question representation, denoted as $q_s = h_m^q$. Similarly, we adopt another LSTM to extract the history features $U^H = (h_0, ..., h_{\ell-1}) \in \mathbb{R}^{d \times \ell}$ at sentence-level, where h_0 is the embedding feature of image caption C.

As questions in a dialog usually have at least one pronoun (*e.g.*, "it", "they", "he"), the agent is required to discover the relevant textual context in the previous history snippets. We employ a question-conditioned attention to aggregate the history context clues, which aims at tackling the textual co-reference. The whole process is formulated as follows:

$$\begin{cases} z_h = tanh((W_q q_s)\mathbb{1}^\top + W_h U^H); \\ \alpha_h = softmax(P_h z_h); \\ u = \sum_{j=0}^{\ell-1} \alpha_{h,j} U_j^H, \end{cases} \tag{1}$$

where $W_q \in \mathbb{R}^{d\times d}$, $W_h \in \mathbb{R}^{d\times d}$ and $P_h \in \mathbb{R}^{1\times d}$ are learnable parameters, $\mathbb{1} \in \mathbb{R}^{1\times \ell}$ is a vector with all elements set to 1, and $\alpha_{h,j}$ and U_j^H are respective the j-th element of α_h and U^H. $u \in \mathbb{R}^{d\times 1}$ denotes the textual context and is further used to construct the context-aware graph.

3.1.2 Graph Representation.

As visual dialog is an on-going conversation, the relations between different objects in the image frequently dynamically vary according to the conversational context. In order to exactly infer the potential relationship, we build a context-aware graph, which takes both visual and textual context semantics into account. The graph structure (relation between objects) will be later iteratively inferred via an adaptive top-K message passing mechanism in Sec. 3.2. Here we construct a graph $G = \{\mathcal{N}, \mathcal{E}\}$, where the i-th node $\mathcal{N}_i$ denotes a joint context feature, corresponding to the i-th visual object feature v_i and its related context feature c_i; the directed edge $\mathcal{E}_{j\rightarrow i}$ represents the relational dependency from node $\mathcal{N}_j$ to node $\mathcal{N}_i$ $(i, j \in [1, n])$. Considering the iterative step t, the graph is denoted as $G^{(t)} = \{\mathcal{N}^{(t)}, \mathcal{E}^{(t)}\}$. There are two cases of $\mathcal{N}^{(t)}$:

$$\begin{cases} \mathcal{N}^{(t)} = (\mathcal{N}_1^{(t)}, ..., \mathcal{N}_n^{(t)}); \\ \mathcal{N}_i^{(t=1)} = [v_i; u]; \;\; \mathcal{N}_i^{(t>1)} = [v_i; c_i^{(t)}], \end{cases} \tag{2}$$

where the iterative step t is initialized as $t = 1$, $[;]$ is the concatenation operation, u is the textual context calculated by Eq. (1) and $\mathcal{N}^{(t)} \in \mathbb{R}^{2d\times n}$. For node $\mathcal{N}_i^{(t)}$ at the iterative step t, the object feature v_i is fixed, and we focus on the context representation $c_i^{(t)}$ learning.

3.2 Iterative Dynamic Directed-Graph Inference

Visual dialog contains implicit relationships between the image, question and history. In order to capture exact context-aware semantics, we exploit the question command $q_w^{(t)}$ (where iterative step $t = 1, ..., T$) to observe new message passing in the graph structure, which is similar to imitating humans checking out the implicit clues multiple times. It is worth noting that our solution, the "**dynamic directed-graph inference**" process, considers flexible, effective, and relevant message propagation. As shown in Fig. 5, at each inference iteration, the context-aware graph is updated through two steps: (1) adjacent correlation matrix learning. Under the instruction of the current question command, each node adaptively selects the top-K most relevant nodes as its neighbors based on an adjacent correlation matrix; (2) top-K message passing. To capture the latent correlation clues for the relational graph learning, each node receives messages from its top-K neighbors and aggregates these messages to update its context feature.

3.2.1 Question-conditioned Relevance Feedback via Adjacent Correlation Matrix Learning.

To infer a correct answer, we essentially have to discover accurate semantics of question Q. At each iterative step t, reviewing different words in Q is helpful to locate keywords. Based on the encoded sequence of question U^Q = $(h_1^q, ..., h_m^q)$, we employ the weighted aggregation to obtain the word attention distribution $\alpha_q^{(t)}$. Then, the word embedding sequence $\mathcal{W}^Q = (w_1, ..., w_m)$ is jointly aggregated to get the question feature at word-level. The whole process is formulated as follows:

$$\begin{cases} z_q^{(t)} = L2Norm(f_q^{(t)}(U^Q)); \\ \alpha_q^{(t)} = softmax(P_q^{(t)} z_q^{(t)}); \\ q_w^{(t)} = \sum_{j=1}^{m} \alpha_{q,j}^{(t)} w_j, \end{cases} \tag{3}$$

where $f_q^{(t)}(.)$ denotes a two-layer MLP, and $P_q^{(t)} \in \mathbb{R}^{1\times d}$ is a learnable parameter. $f_q^{(t)}(.)$ and $P_q^{(t)}$ are learned at t-th graph inference step. $q_w^{(t)} \in \mathbb{R}^{d_w\times 1}$ denotes the t-th question command.

After obtaining the question command $q_w^{(t)}$, we measure the correlation among different nodes in the graph. We design an adjacency correlation matrix of the graph $G^{(t)}$ as $A^{(t)} \in \mathbb{R}^{n\times n}$, in which each value $A_{l\rightarrow i}^{(t)}$ represents the connection weight of the edge $\mathcal{E}_{l\rightarrow i}^{(t)}$. We learn the correlation matrix $A^{(t)}$ by computing the similarities of each pair of nodes in $\mathcal{N}^{(t)}$ under question command $q_w^{(t)}$.

$$A^{(t)} = (W_1\mathcal{N}^{(t)})^\top((W_2\mathcal{N}^{(t)}) \odot (W_3 q_w^{(t)})), \tag{4}$$

where $W_1 \in \mathbb{R}^{d\times 2d}$, $W_2 \in \mathbb{R}^{d\times 2d}$ and $W_3 \in \mathbb{R}^{d\times d_w}$ are learnable parameters. $\odot$ denotes the hadamard product, *i.e.*, element-wise multiplication.

As we all know, in the image each object is related to only a small subset of the other objects, especially rare under the instruction of the question (*i.e.*, sparse relationship). Therefore, each node in the graph is required to connect with the most relevant neighborhood nodes. In order to learn a set of related neighborhoods $S_i^{(t)}$ of each node $\mathcal{N}_i^{(t)}$, $i \in [1, n]$, we adopt a ranking strategy as:

$$S_i^{(t)} = topK(A_i^{(t)}), \tag{5}$$

where $topK$ returns the indices of the K largest values of an input vector, and $A_i^{(t)}$ denotes the i-th row of the adjacent matrix. Note that each node has its independent neighbors $S_i^{(t)}$. As shown in Fig. 4 $(topK, K = 2)$, even the same node can also have different neighbors at each inference step. It means that our solution is a dynamic graph inference process. Our CAG graph is an asymmetric directed-graph.

3.2.2 Relational Graph Learning via Top-K Message Passing.

The current graph structure is relational-aware. The influence of each node can be measured by its K neighborhood nodes. We propagate the relational clues to each node via a message passing mechanism. Taking node $\mathcal{N}_i^{(t)}$ at the t-th step inference as example, it receives the messages from its most K relevant neighborhood nodes $\{\mathcal{N}_j^{(t)}\}$, where $j \in S_i^{(t)}$.

To measure the influences of these neighbors, $B_{j\rightarrow i}^{(t)}$ normalizes the connection weight of $\mathcal{N}_j^{(t)}$ to $\mathcal{N}_i^{(t)}$ with a $softmax$ function. As shown in Eq. (6), in the adjacent correlation matrix $A^{(t)}$, $A_{j\rightarrow i}^{(t)}$ denotes the connection weight of the edge $\mathcal{E}_{j\rightarrow i}^{(t)}$. Under the instruction of the question command $q_w^{(t)}$, $m_{j\to i}^{(t)}$ calculates the incoming message of neighbour $\mathcal{N}_j^{(t)}$ to $\mathcal{N}_i^{(t)}$. At last, $\mathcal{N}_i^{(t)}$ sums up all the incoming messages to get the final message feature $M_i^{(t)}$. The whole process is formulated as follows:

$$\begin{cases} [B_{j\to i}^{(t)}] = \underset{j}{softmax}([A_{j\to i}^{(t)}]), \quad j \in S_i^{(t)}; \\ m_{j\to i}^{(t)} = (W_4 \mathcal{N}_j^{(t)}) \odot (W_5 q_w^{(t)}); \\ M_i^{(t)} = \sum_j B_{j\to i}^{(t)} m_{j\to i}^{(t)}, \end{cases} \tag{6}$$

where $W_4 \in \mathbb{R}^{d\times 2d}$ and $W_5 \in \mathbb{R}^{d\times d_w}$ are learnable parameters. $M_i^{(t)} \in \mathbb{R}^{d\times 1}$ denotes the summarized message to $\mathcal{N}_i^{(t)}$, and the node representation $\mathcal{N}_i^{(t)}$ is then updated to $\mathcal{N}_i^{(t+1)}$:

$$\begin{cases} c_i^{(t+1)} = W_6[c_i^{(t)}; M_i^{(t)}]; \\ \mathcal{N}_i^{(t+1)} = [v_i; c_i^{(t+1)}], \end{cases} \tag{7}$$

where $W_6 \in \mathbb{R}^{d\times 2d}$. In Eqs. 4 $\sim$ 7, W_1 $\sim$ W_6 are learnable parameters that are shared for each iteration. After performing T-steps message passing iterations, the final graph node representation is denoted as $\mathcal{N}^{(T+1)}$.

3.3 Graph Attention Embedding

Up to now, the context clues on each node in the graph $\mathcal{N}^{(T+1)}$ not only correspond to visual and textual features, but also involve iteratively relational context learning. As the majority of questions usually pay attention to a part of objects in the image scene and history snippets, we apply a question-conditioned graph attention mechanism to selectively attend the graph nodes. The whole graph attention is learned as follows:

$$\begin{cases} z_g = tanh((W_{g_1} q_s)\mathbb{1}^\top + W_{g_2}\mathcal{N}^{(T+1)}); \\ \alpha_g = softmax(P_g z_g); \\ e_g = \sum_{j=1}^{n} \alpha_{g,j} \mathcal{N}_j^{(T+1)}, \end{cases} \tag{8}$$

where $W_{g1} \in \mathbb{R}^{d\times d}$ and $W_{g2} \in \mathbb{R}^{d\times 2d}$ are learnable parameters. $e_g \in \mathbb{R}^{2d\times 1}$ denotes the attended graph embedding.

Finally, we fuse it with textual context u and question feature q_s to output the multi-modal embedding $\widetilde{e}$:

$$\widetilde{e} = tanh(W_e[e_g; u; q_s]). \tag{9}$$

The output embedding $\widetilde{e}$ is then fed into a softmax decoder to sort the candidate answers in A, and choose the answer with the highest probability as the final prediction. The loss function of answer decoding is formulated as follows:

$$L_{n-pair} = log\left(1 + \sum_{i=1}^{N} exp(\widetilde{e}^\top f(a_i^-) - \widetilde{e}^\top f(a^{gt}))\right). \tag{10}$$

where L_{n-pair} is a metric-learning multi-class N-pair loss [33]. $f(.)$ is an self-attention based LSTM encoder for the answer, a^{gt} is the ground-truth answer, and a_i^- represents incorrect options.

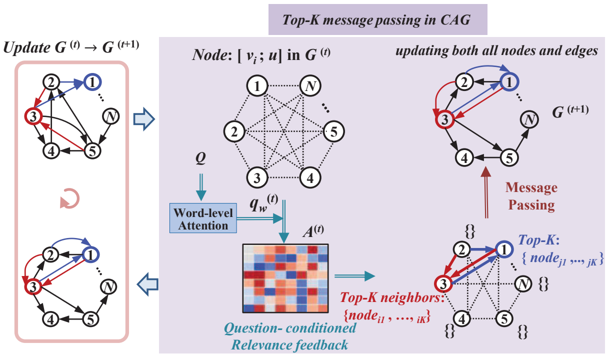

Top-K message passing in CAG
Update G (t) → G (t+1)
Node: [vi ; u] in G (t)
updating both all nodes and edges
G (t+1)
1
2
3
4
5
N
Q
Word-level Attention
qw(t)
A(t)
Message Passing
Top-K: { nodej1 ,..., jK }
Top-K neighbors: {nodei1 , …, iK}
{}
Question- conditioned Relevance feedback

Fig. 5. Adaptive Top-K Message Passing Unit.

4 Optimization: Visual-Aware Knowledge Distillation

After learning the relational graph, here we restrain the historical noise (as shown in Fig. 3) to obtain a more robust graph. We design a visual-aware distillation scheme. The usage of knowledge distillation is different from classical methods [30], [48]. In classical distillation, a teacher model is reliable with superior performance, which instructs the student model how to learn useful knowledge. In our work, we target to learn and retain the consistency of common visual awareness under both teacher (Img-Only CAG) and student (Joint CAG) graphs models.

Concretely, our proposed distillation scheme imposes soft regularization learned from teacher model onto student model, not only alleviates over-learning of student model under hard constraints, but also makes the student model more robust with rich visual context awareness. This distillation scheme riches the to-be-learned knowledge in the graph learning process, where the entropy and consistency of soft targets are evaluated. It is an accessible way to obtain soft labels with probability distribution, which provides more semantic clues than hard constraints with one-hot ground-truth labels in our task.

4.1 Teacher-Image & Student-Joint CAG Models

Teacher Graph: Img-Only CAG. In this graph, we learn pure visual relation merely under the guidance of the question, even sometimes with ambiguous pronoun words. Vision is a very important factor in an image-centric conversation. In this case, we merely learn the visual relation in the dialog without historical clues. We randomly initialize $c_i^{(1)}$ in Eq. 2: $\mathcal{N}_i^{(t=1)} = [v_i; c_i^{(1)}]$; thus $c_i^{(1)}$ no longer involves the history clues.

Student Graph: Joint CAG. The Joint CAG graph denotes the entire graph CAG introduced in Sec. 3, where $c_i^{(1)}$ = u. Different from that the absence of history in the graph learning of Img-Only CAG, here we explicitly introduce the historical cues in the whole graph inference process. To avoid over-learning history snippets, we have to balance the Img-Only and Joint CAG models.

4.2 Distillation Scheme

To ensure a robust graph inference, we leverage the pure visual relation (visual-aware knowledge) at object-level learned by the Img-Only CAG model to regularize the Joint CAG model. Note that the parameters of the Image-only and Joint CAG models are not shared in our framework. Both image-Only and Joint CAGs are trained by the loss optimization L_{n-pair} (Eq. 10) in advance. Historical clues cannot be completely neglected in the relational reasoning; hence, we select the Joint CAG model as a backbone and optimize it by knowledge distillation (Eq. 13). During testing, only the Joint model is used.

The essence can be thought of learning implicit visual relational consistency of Img-Only and Joint CAG graphs. To be specific, we attempt to find a balance on the proposed CAG - keeping the generalization ability of teacher (Img-Only CAG) and discrimination ability of student (Joint CAG) on context-awareness. The graph learning is essentially constrained by the probability distributions of candidate answers; we conduct the distillation scheme by applying soft regularization on answer logits. In this way, we learn a robust CAG model in a trainable mode, rather than directly applying score fusion as ensemble models do [25], [54], [55].

We minimize the KL divergence [30], [31] between the answer prediction probability distributions from both Img-Only and Joint CAG models. The loss function is formulated as follows:

$$L_{KL} = -\sum_{i=1}^{100} P_{J,i} log(\frac{P_{I,i}}{P_{J,i}}), \tag{11}$$

where $P_{J,i} = softmax(\widetilde{e}^{\top} f(a_i))$ is the probability (softmax logits) of the candidate answer a_i obtained by Joint CAG, and $P_{I,i}$ is the probability of a_i generated by Img-only CAG.

Besides, we adopt a standard cross-entropy loss L_{CE} to measure the entropy of the answers generated by the Joint CAG model :

$$L_{CE} = -\sum_{i=1}^{100} y_i log(P_{J,i}), \tag{12}$$

where y is a one-hot encoded vector of the ground-truth answer.

The distillation loss function consists of these two parts:

$$L_{distill} = L_{CE} + \lambda L_{KL}, \tag{13}$$

where λ is a trade-off hyper-parameter.

Until now, there are different model settings of CAG. For the sake of convenience in the following representations, we abbreviate the Joint CAG model without knowledge distillation as CAG, and with knowledge distillation as CAG-Distill. Namely, **"CAG = Joint CAG w/o Knowledge Distillation"** and **"CAG-Distill = Joint CAG w Knowledge Distillation"**. More details of the training settings are explained in Sec. 5.1.3.

5 Experiments

5.1 Experiment Setup

5.1.1 Datasets.

Experiments are conducted on benchmark datasets VisDial v0.9 and v1.0 [21]. In VisDial v0.9, the dialog consists of 10-round QA pairs for each image. VisDial v0.9 contains 83k and 40k dialogs on COCO-train and COCO-val images [56] respectively, totally 1.2M QA pairs. VisDial v1.0 is an extension of VisDial v0.9, which adds additional 10k dialogs on Flickr images. The new train, validation, and test splits contains 123k, 2k and 8k dialogs, respectively. It is worth noting that in the test split of VisDial v1.0, each dialog has flexible m rounds of QA pairs, where m is in the range of 1 to 10. Besides, in VisDial v1.0, validation and test datasets are annotated with relevance scores by four human [1], *e.g.*, some semantically identical candidate answers, "i can't tell" and "i cannot tell" with scores 1.0 and 0.8. The relevance scores are used to evaluate whether the model can predict all the relevant answers with high ranks.

5.1.2 Evaluation Metrics

Following [21], the answer accuracy is evaluated by retrieving the ground-truth answer from a list of 100 option answers. We adopt the following retrieval metrics: (1) average rank of the ground-truth answer (**Mean**), (2) recall rate of the ground-truth answer in top-k ranked option answers (**R@k**), (3) mean reciprocal rank of the ground-truth answer (**MRR**), and (4) normalized discounted cumulative gain (**NDCG**). VisDial v1.0 introduces a new retrieval metric **NDCG** to further evaluate the models generalization ability. The questions usually can be responsed with more than one correct options in the candidate set, such as "yes" and "yes, it is". In this situation, NDCG is invariant to the order of options with identical relevance and to the order of options outside of the top k. The metric is given by:

$$\begin{cases} DCG@k = \sum_{i=1}^{k} \frac{relevance_i}{log_2(i+1)}; \\ NDCG@k = \frac{DCG@k \ for \ predicted \ ranking}{DCG@k \ for \ ideal \ ranking}, \end{cases} \tag{14}$$

where $relevance_i$ is relevance score that annotated by people who think the option answer is relevant to the question, and k is the number of option answers with non-zero relevance scores. For these metrics, a higher score is better for MRR, R@k, and NDCG, while a lower score is better for Mean.

5.1.3 Implementation Details.

Language Processing. For pre-processing text data Q and H, we lowercase all words and remove contractions in questions and answers, and use the Python NLTK toolkit to tokenize the sentences in the datasets. Next, we retain the words that occur at least 4 times in the training split, resulting in a vocabulary of 9,793 words for VisDial v0.9 dataset and 11,319 words for VisDial v1.0 dataset. And the captions, questions, and answers are padded or truncated to 40, 20 and 20, respectively. Each word in the dialog is embedded into a 300-dim vector by the GloVe embedding initialization [57]. We set all the LSTMs in the model with 1-layer and 512 hidden states.

Training Details. Our model training process can be divided into two stages. In the first training stage, We pretrain the Img-Only and Joint CAG models with multi-class N-pair loss (Eq. 10) [33], respectively. For both models, we adopt Adam optimizer [58] and initialize the learning rate

[1] https://visualdialog.org

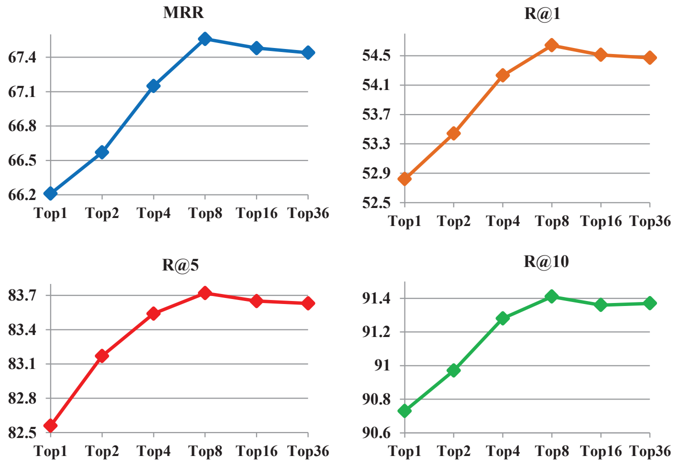


Fig. 6. Performance comparison of the neighborhood number K on VisDial val v0.9

with 4×10^{-4}. The learning rate is multiplied by 0.5 after every 10 epochs. In the second training stage, we first use the pre-trained Img-Only model to generate soft labels; then, we use them to further train the Joint CAG model with both KL divergence loss and cross-entropy loss (Eq. 13). The hyper-parameter λ in the loss function is set to be 10. The learning rate is initialized with 5×10^{-5} and multiplied by 0.5 after every 4 epochs. In both training stages, we apply Dropout [59] with ratio 0.3 for attention layers and the last fusion layer. All of our experiments are implemented on the platform of Pytorch.

5.2 Ablation Study of CAG

5.2.1 Empirical Parameters

Neighborhood Number (K)**.** We test different neighborhood numbers $K \in \{1, 2, 4, 8, 16, 36\}$. As shown in Fig. 6, $K = 8$ is an optimal parameter setting. Performances drop significantly for $K < 8$. It means that if the selected neighborhood nodes are insufficient, the relational messages can not be fully propagated. While setting the neighborhood number $K > 8$, the node receiving redundant irrelevant messages from neighbors can disturb the reasoning ability of the model. Thus, we set the neighborhood number $K = 8$ in the following experiments.

Iteration Steps (T)**.** T indicates the number of relational reasoning steps to arrive the answer. We test different steps T to analysis the influence of iterative inferences. As shown in Table 1, the performance of CAG is gradually improved with the increasing T. We have the best performance at $T = 3$, lifting R@1 from 53.25 (T = 1) to 54.64. The proposed iterative graph inference is effective. Visualization results in Fig. 9 further validate this result. When $T > 3$, the performance drops slightly. It means that if the relationships in the graph have been fully inferred, further inference does not help. The questions in the VisDial datasets are collected from relative simple free-form human dialogue. The setting of T = 3 already performs well. In the following experiment, we set $T = 3$.

TABLE 1
Ablation studies of different iteration steps T and the main components on VisDial val v0.9.

Model	Step T	Mean↓	MRR↑	R@1↑	R@5↑	R@10↑
CAG	T = 1	4.02	66.32	53.25	82.54	90.55
	T = 2	3.91	66.93	53.76	83.11	90.96
	T = 3	**3.75**	**67.56**	**54.64**	**83.72**	**91.48**
	T = 4	3.83	67.28	54.11	83.46	91.17
	T = 5	3.80	67.55	54.63	83.48	91.14
CAG w/o Infer	-	4.11	65.73	52.56	82.38	90.36
CAG w/o u	T = 3	4.19	65.26	51.83	81.55	90.21
CAG w/o Q-att	T = 3	3.91	66.70	53.74	82.75	90.89
CAG w/o G-att	T = 3	3.86	66.98	53.99	83.08	91.04
CAG	T = 3	**3.75**	**67.56**	**54.64**	**83.72**	**91.48**

Note that "CAG = Joint CAG".

5.2.2 Main Components

A few variants are proposed for ablation study. **CAG w/o Infer** denotes that CAG removes the whole dynamic directed-graph inference in Sec. 3.2. It means that all the nodes and edges in the graph will not be updated and inferred. **CAG w/o *u*** denotes that CAG without textual-history context u, where the whole graph merely describes visual context clues. **CAG w/o Q-att** denotes CAG without word-level attention on question Q. **CAG w/o G-att** removes the graph attention module. All the node representations are average pooled in the graph embedding phase.

As shown in Table 1, compared with **CAG**, **CAG w/o Infer** drops MRR significantly from 67.56 to 65.73. It indicates that the graph inference effectively performs well for relational reasoning. Learning the implicit relations between the nodes is helpful to predict the final answer. **CAG w/o *u*** drops R@1 significantly from 54.64 to 51.83. It indicates that the joint visual-textual context learning is necessary. Relations between the nodes can not be fully inferred without textual clues u. **CAG w/o Q-att**, which replaces question commands $\{q_w^{(t)}\}$ with the sentence-level feature q_s, drops R@1 from 54.64 to 53.74. It also can be explained. Figs. 9∼10 demonstrate that the attentive words always vary during the inference process. It usually firstly identifies the target in the question, then focuses on related objects, and finally observes attributes and relations in both visual and textual context clues to infer the answer. **CAG w/o G-att**, which removes the final graph attention module, drops R@1 from 54.64 to 53.99. Although each node involves relational reasoning, not all the nodes are relevant to the current question. Thus, paying attention to relevant nodes in the relational graph is helpful to infer an exact answer.

5.3 Ablation Study of Knowledged Distillation

In this subsection, we evaluate each impact of and the complementarity between Img-Only and Joint CAG models. Then, we introduce the Img-Only model to distill the knowledge into the Joint model.

5.3.1 Role of the Dialog History

In the Visual Dialog evaluation metrics, MRR is measured based on the predicted ranking of ground-truth answers, while NDCG measures the ranking of all answers that have similar meanings to ground-truth. If the MRR reflects the

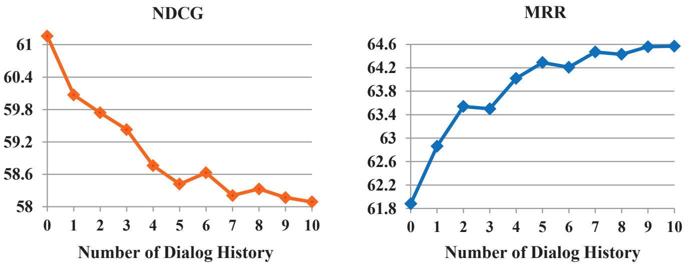


Fig. 7. Performance comparison of CAG randomly involving different amounts of history on the VisDial v1.0 validation dataset.

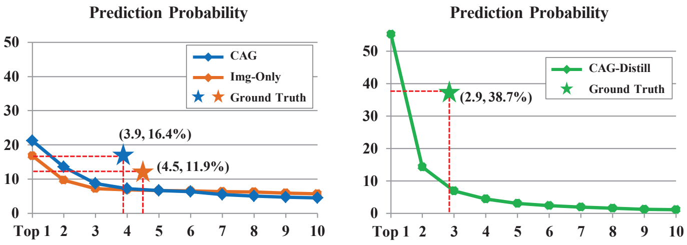


Fig. 8. Probability distribution of the top-10 predicted answers on the VisDial v1.0 validation dataset, where each star marks the average rank order and score of the ground-truth answer. The sharper probability curve has a much more powerful discrimination capability. CAG-Distill performs significant improvement.

preciseness of a model, then NDCG reflects the generalization ability. As shown in Fig. 7, as the amount of dialog history increases, the NDCG performance decreases significantly, and the best NDCG value 61.16 is achieved when CAG does not introduce any history information (*i.e.*, Img-Only model). Does it mean the Img-Only CAG can reason the answer more correctly? The answer may be no. The MRR metric is the worst at the Img-Only setting, which gradually improves with the increase of the amount of history, and the best MRR value 64.57 is achieved when CAG considers all the history (*i.e.*, the Joint CAG model). R@1 $\sim$ R@10 values in Table 5 also validate this conclusion.

In this paper, we attempt to keep a good generalization ability of the model, and also retain discrimination ability with high preciseness. Indeed Img-only model can generalize the visual awareness ability of the model, but its discrimination ability seems to be unconfident with low predicted probability scores. As shown in Fig. 8, as for the top-10 predicted ranking list on the VisDial v1.0 validation dataset, the predicted probability curve of Img-Only is the smoothest, which indicates Img-Only CAG outputs more similar probability scores, with a slight advantage of distinguishing the correct answers over incorrect ones. By contrast, CAG-Distill in Fig. 8 performs obviously superior discrimination ability than others. Knowledge distillation between Img-Only and Joint CAG models takes effect. CAG-Distill performs an excellent robust power.

Another argument may be that whether the proposed model prefers the high-frequency words in history? We have collected a specific subset form VisDial v1.0 val, in which questions contain high-frequency words "yes" or "no". As the statistics of VisDial v1.0 val, it contains 2,064 dialog samples with 20,640 questions. Among these questions, we selected the questions whose ground-truth answer is "yes" or "no", giving us with 8,060 questions. We then further removed the samples that do not include "yes" or "no" in the dialog history and the samples in which the word frequencies of "yes" and "no" are consistent. Finally, we acquired 6,778 questions. We call this subset of VisDial v1.0 val as VisDial v1.0 (val-yn), which has obviously semantic inclination to high-frequency words "yes" or "no".

TABLE 2
Answer accuracy comparison on VisDial v1.0 (val-yn).

Model	All	'Yes'	'No'
	6778	3800	2978
Count-Words	41.66%	34.55%	50.73%
CAG	74.53%	74.07%	75.11%
CAG-Distill	77.61%	77.26%	78.04%

TABLE 3
Performance of the ensemble models between Image-Only model and Joint model on VisDial val v1.0.

Metrics	Img-Only	Joint	Distill	$Ens_{gt_{min}}$	$Ens_{gt_{max}}$
NDCG	61.16	58.09	59.26	55.22	64.02
MRR	61.88	64.57	65.38	56.88	69.57

Img-Only: the graph inference of CAG with input (Q, I), Joint: the graph inference of CAG with input (Q, H, I). Distill: the Joint CAG model with distillation optimization. $Ens_{gt_{min}}$: for each question, the worse prediction result of the ground-truth answer in the two models are served as the final result; $Ens_{gt_{max}}$: for each question, the better prediction result of the ground-truth answer in the two models are served as the final result.

As shown in Table 2, merely counting the high-frequency words "yes" or "no" in history to predict the answer (**Count-Words**) does not work well with 41.66% accuracy at the *ALL* metric. In contrast, **CAG** and **CAG-Distill** utilize image I, question Q and history H to infer the answer. Both achieved better performances, showing accuracies of 74.54% and 77.61% respectively, at *ALL* metric. Fig. 15 visualizes some positive examples inhibiting the negative influences of high-frequency words in history. In a nutshell, on the VisDial v1.0 (val-yn) dataset, CAG and CAG-Distill models exhibit a much better semantic understanding capacity compared with "Count-Word".

5.3.2 Model Complementary

An ideal model should have balanced performances over all the metrics rather than having higher scores only for a certain metric. We attempt to find a balance on the proposed CAG with both good preciseness and generalization. Can the advantages of both models be combined? We test the model complementary on the VisDial val v1.0 dataset. As shown in Table 3, we adopt two simple ensemble strategies on Img-Only and Joint CAG models with the prediction rankings. (1) Minimum ensemble strategy ($Ens_{gt_{min}}$)for each question, the worse prediction result of the ground-truth answer in the two models are served as the final result; (2) Maximum ensemble strategy ($Ens_{gt_{max}}$: for each question, the better prediction result of the ground-truth answer in the two models are served as the final result. To some extent, $Ens_{gt_{max}}$ reflects a trend of upper limit of CAG under the ground-truth verification. Once the Img-Only and Joint CAG models are effectively combined, the

TABLE 4
Main evaluation of discriminative models on both VisDial v0.9 and v1.0 datasets.

Model	VisDial v0.9 (val)					VisDial v1.0 (test-std)					
	Mean↓	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓	NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑
Fusion-based Models											
LF [21]	5.78	58.07	43.82	74.68	84.07	5.95	45.31	55.42	40.95	72.45	82.83
HRE [21]	5.72	58.46	44.67	74.50	84.22	6.41	45.46	54.16	39.93	70.45	81.50
Attention-based Models											
HREA [21]	5.66	58.68	44.82	74.81	84.36	-	-	-	-	-	-
MN [21]	5.46	59.65	45.55	76.22	85.37	5.92	47.50	55.49	40.98	72.30	83.30
RA-Net [60]	4.95	62.27	48.63	78.62	87.49	-	-	-	-	-	-
HCIAE [33]	4.81	62.22	48.48	78.75	87.59	-	-	-	-	-	-
AMEM [36]	4.86	62.27	48.53	78.66	87.43	-	-	-	-	-	-
CoAtt [34]	4.47	63.98	50.29	80.71	88.81	-	-	-	-	-	-
CorefNMN [37]	4.45	64.10	50.92	80.18	88.81	4.40	54.70	61.50	47.55	78.10	88.80
DVAN [35]	3.93	66.67	53.62	82.85	90.72	4.36	54.70	62.58	48.90	79.35	89.03
RVA [39]	3.93	66.34	52.71	82.97	90.73	4.18	55.59	63.03	49.03	80.40	89.83
Synergistic [54]	-	-	-	-	-	4.17	57.32	62.20	47.90	80.43	89.95
RAA-Net [26]	3.89	66.83	53.80	82.99	90.86	4.35	55.42	62.86	49.05	79.65	88.85
DAN [55]	4.04	66.38	53.33	82.42	90.38	4.30	57.59	63.20	49.63	79.75	89.35
HACAN [61]	3.97	67.92	54.76	83.03	90.68	4.20	57.17	64.22	50.88	80.63	89.45
ReDAN‡ [25]	-	-	-	-	-	6.63	64.47	53.73	42.45	64.68	75.68
MCA-I-H† [27]	-	-	-	-	-	8.89	**72.47**	37.68	20.67	56.67	72.12
Graph-based Models											
GNN [23]	4.57	62.85	48.95	79.65	88.36	4.57	52.82	61.37	47.33	77.98	87.83
FGA w/o Ans [24]	4.63	62.94	49.35	79.31	88.10	-	-	-	-	-	-
FGA [24]	4.35	65.25	51.43	82.08	89.56	4.51	52.10	63.70	49.58	**80.97**	88.55
DualVD [29]	4.17	62.94	48.64	80.89	89.94	4.11	56.32	63.23	49.25	80.23	89.70
CAG (Ours)	3.75	67.56	54.64	83.72	91.48	4.11	56.64	63.49	49.85	80.63	90.15
CAG-Distill (Ours)	**3.71**	**68.06**	**55.26**	**83.98**	**91.58**	**4.05**	57.77	**64.62**	**51.28**	80.58	**90.23**

‡denotes ensemble model and †indicates fine-tuning on dense annotations. Bold font and underline respectively denote the 1st and 2nd best performances.

preciseness and generalization ability of the model can be further improved, such as CAG-Distill. More experimental results of CAG-Distill are introduced in the following parts.

5.3.3 Distillation Analysis

To enhance the generalization ability of the CAG, we treat the Img-Only model as a teacher model that transfers the learned knowledge (soft probability scores) to regularize the Joint model. Observing Table 5, the distilled CAG models significantly outperform "Joint", *e.g.,* lifting NDCG from 58.09 to 59.26 and MRR from 64.57 to 65.38. These results validate the effectiveness of knowledge distillation. **CAG-Distill w/o KL** denotes that the model does not calculate the KL divergence loss (*i.e.,* soft labels - answer softmax logits - are not accessible). **CAG-Distill w/o KL** achieves slight improvement compared with "Joint" CAG.

Furthermore, there are two typical types of distillation - distilling knowledge at feature embedding level or answer probability level. [52] belongs to the former, and ours involves the latter. In other words, we apply the distillation scheme in [52] to the graph learning of Img-Only and CAG. The total distillation loss is formulated as follows:

$$\begin{cases} L_{cos} = 1 - cos(\widetilde{e}_I, \widetilde{e}_J); \\ L_{distill} = L_{CE} + \lambda L_{cos}, \end{cases} \tag{15}$$

TABLE 5
Performance of the knowledge distillation model on VisDial val v1.0.

Model	Mean↓	NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑
Img-Only	4.42	**61.16**	61.88	47.82	79.16	88.84
Joint	4.01	58.09	64.57	51.12	81.43	90.30
Distill w/o KL	4.04	58.29	64.90	51.49	81.49	90.39
GraEm-Distill [52]	3.97	58.51	65.08	51.62	81.88	90.59
Ans-Distill (Ours)	**3.94**	59.26	**65.38**	**51.94**	**82.14**	**90.75**

Note that "CAG-Distill = CAG-Ans-Distill".

where, $\widetilde{e}_I$ denotes the multi-modal embedding of Img-Only (*i.e.,* final output of graph embedding as formulated in Eq. 9), and $\widetilde{e}_J$ is the final output multi-modal embedding of (Joint) CAG. λ is set to 0.5 for L_{cos}.

To differentiate these two distillation schemes, we abbreviate our soft regularizer with answer logits as **Ans-Distill (ours)** and the regularizer [52] with multi-modal embedding vectors after the graph inference as **GraEm-Distill**. As shown in Table 5, both **GraEm-Distill** and **Ans-Distill** are effective. Compared with CAG without distillation, **GraEm-Distill** performs better with the Mean of from 4.01 to 3.97 and MRR increase from 64.57 to 65.08. However, **Ans-Distill (ours)** performs the best (*i.e.,* Mean of 3.94 and MRR of 65.38). We owe this to the nature of the task. Knowledge distillation at the answer probability level is tractable for the answer prediction. Distilling the feature embedding may introduce noise between the teacher model (Img-Only) and the student model (CAG). Knowledge distillation at the answer probability level is tractable for the answer prediction.

5.4 Comparison Results

In this section, we compare the proposed CAG with the existing state-of-the-art approaches as follows: (1) **Fusion-based Models** (LF [21] and HRE [21]); (2) **Attention-based Models** (HREA [21], MN [21], HCIAE [33], AMEM [36], CoAtt [34], CorefNMN [37], DVAN [35], RVA [39], Synergistic [54], DAN [55], RAA-Net [26], and HACAN [61]); and (3) **Graph-based Methods** (GNN [23], FGA [24], and DualVD [29]).

5.4.1 Results on VisDial v0.9.

As shown in Table 4, CAG consistently outperforms most of methods. Compared with fusion-based models **LF** [21] and **HRE** [21], the R@1 performance of our **CAG** is significantly improved, lifting each other by 10.8% and 9.9%. For attention-based models, compared to a previous state-of-the-art model **DAN** [55], our model outperforms it at all evaluation metrics. **HACAN** [61] reports the recent best results. It first pre-trains with N-pair loss, and then uses the wrong answers to "tamper" the truth-history for data augment. Finally, the truth- and fake-history are used to fine-tune its model via reinforcement learning. Without the fine-tuning tactic, CAG still outperforms **HACAN** on Mean, R@5, and R@10. CAG-Distill outperforms HACAN in all evaluation metrics.

Here, we mainly compare our method with the graph-based models. **GNN** [23] constructs a graph that only exploring the dependencies between the textual-history. In contrast, our **CAG** builds a graph over both visual-objects and textual-history contexts. Compared with **GNN**, our model achieves 5.7% improvements on the R@1 metric. **FGA** [24] is the previous state-of-the-art graph-based method for visual dialog, which treats the candidate answer embedding feature A as new context clue and introduces it into the multi-modal encoding training. This operation improves their results a lot (FGA w/o Ans $vs.$ FGA). Without candidate answer embedding, our model still performs better results, lifting R@1 from 51.43 to 54.64, and decreasing the Mean from 4.35 to 3.75. **DualVD** [29] is a recently proposed method, which employs the graph to model the relations among the image objects and uses the question-guided graph convolution for relational reasoning. In addition, DualVD introduces dense image captions to get external knowledge for better semantic learning. Without external knowledge, our CAG still outperforms it in all evaluation metrics. In our solution, we merely utilize fine-grained visual-textual semantics that are helpful for answer inferring. After fine-tuning by the knowledge distillation tactic, the performance of CAG has been further improved, lifting MRR from 67.56 to 68.06, and R@1 from 54.64 to 55.26.

Test with VGG Features. As some existing methods evaluated with VGG features, to be fair, we test our model with VGG features too. Table 6 shows that our **CAG-VGG** outperforms previous methods that only utilize VGG features. Compared to **CAG-VGG**, **CAG** and **CAG-Distill** gets a significant performance boost. It indicates the object-region features provide richer visual semantics than VGG features.

TABLE 6
Performance comparison on VisDial val v0.9 with VGG features. Our model with VGG features is denoted as CAG-VGG.

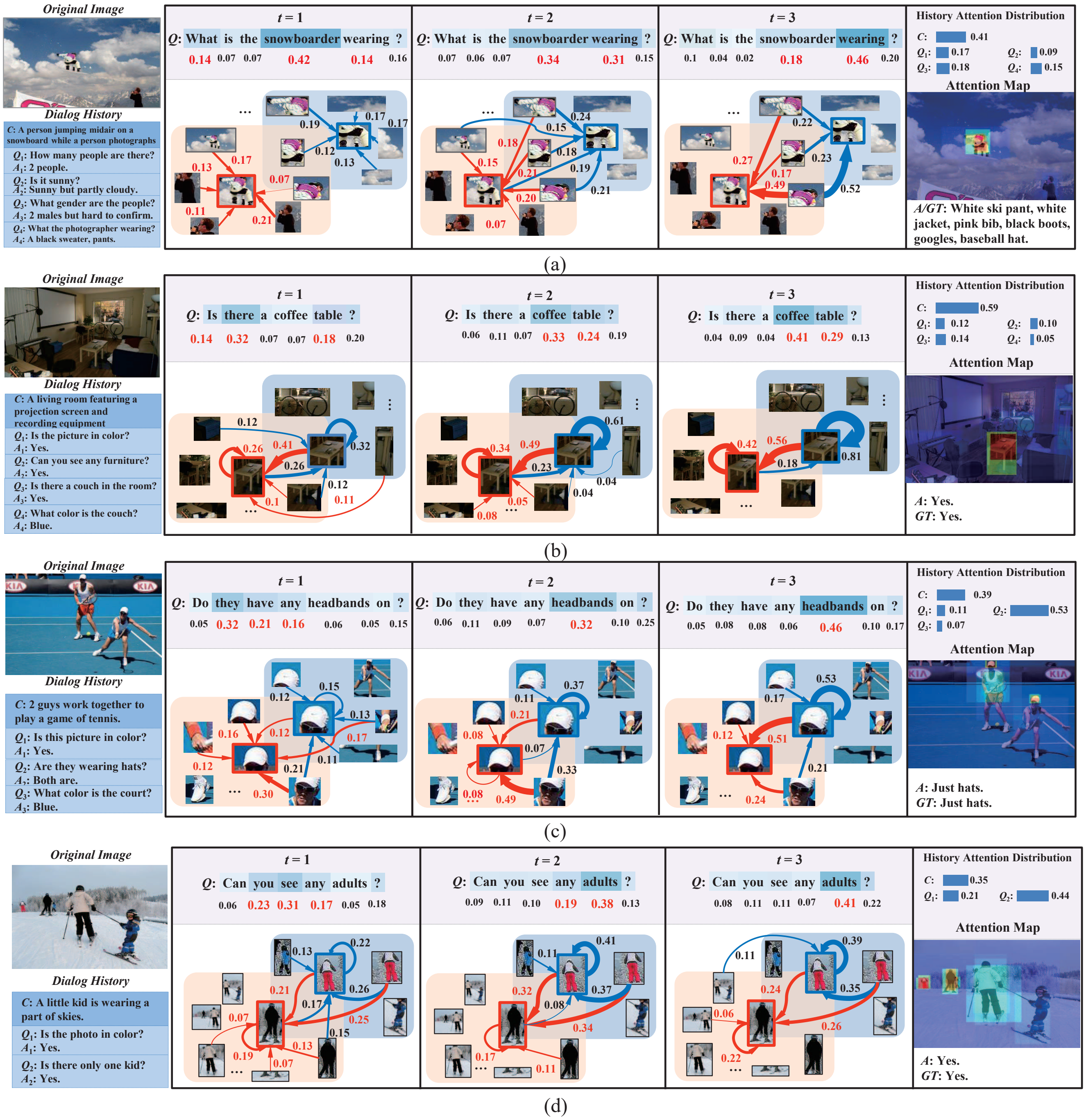
Model	Mean↓	MRR↑	R@1↑	R@5↑	R@10↑
Attention-based Models					
HCIAE [33]	4.81	62.22	48.48	78.75	87.59
AMEM [36]	4.86	62.27	48.53	78.66	87.43
RA-Net [60]	4.95	62.27	48.63	78.62	87.49
CoAtt [34]	4.47	63.98	50.29	80.71	88.81
DVAN [35]	4.38	63.81	50.09	80.58	89.03
RvA [39]	4.22	64.36	50.40	81.36	89.59
HACAN-VGG [61]	4.32	64.51	50.72	81.18	89.23
HACAN [61]	3.97	67.92	54.76	83.03	90.68
Graph-based Models					
GNN [23]	4.57	62.85	48.95	79.65	88.36
FGA w/o Ans [24]	4.63	62.94	49.35	79.31	88.10
CAG-VGG (Ours)	4.13	64.91	51.45	81.60	90.02
CAG-VGG-Distill (Ours)	4.07	65.35	51.89	81.91	90.11
CAG (Ours)	3.75	67.56	54.64	83.72	91.48
CAG-Distill (Ours)	**3.71**	**68.06**	**55.26**	**83.98**	**91.58**

5.4.2 Results on VisDial v1.0.

A new metric NDCG (Normalized Discounted Cumulative Gain) [62] is proposed to evaluate quantitative semantics, which penalizes low ranking correct answers. Other metrics evaluate the rank of the ground-truth in the candidate answer list. NDCG tackles the issue of more than one plausible answers in the answer set. Compared with the attention-based models, as above mentioned, **HACAN** [61] trains the model twice, **Synergistic** [54] sorts the candidate answers twice. Without resorting or fine-tuning, under end-to-end training, our CAG model still performs better performance on the Mean value. Compared with the graph-based models, our model has greatly improved the NDCG value. CAG outperforms **GNN** [23], **FGA** [24], and **DualVD** [29] by 3.8%, 4.5%, and 1.1%, respectively. This also proves that our graph method can reason out more plausible answers. Furthermore, CAG-Distill not only significantly outperforms existing graph-based models, but also performs better than the existing attention-based models. This proves that transferring the learned knowledge from image-centric visual-awareness to image-history joint awareness can effectively improve the generalization ability and preciseness of the model. In addition, in Sec. 5.5, we provide more intuitive visualization results to explain how CAG implements the reasoning process and display the ranking changes of predicted answers in CAG-Distill.

5.4.3 Discussion on Ensemble Model.

In this paper, we discuss the ensemble model to combine both attention and graph-based models. In previous works, researchers have conducted an ensemble experiment on attention models, such as [25] for Visual Dialog. To our knowledge, no work refers to the ensemble model with both graph and attention-based models. To validate the effectiveness, we experimented on the ensemble model on Visdial v1.0 Val. We have the code of the attention-based



(a)

C: A person jumping midair on a snowboard while a person photographs

Q_1: How many people are there?
A_1: 2 people.

Q_2: Is it sunny?
A_2: Sunny but partly cloudy.

Q_3: What gender are the people?
A_3: 2 males but hard to confirm.

Q_4: What the photographer wearing?
A_4: A black sweater, pants.

A/GT: White ski pant, white jacket, pink bib, black boots, googles, baseball hat.

(b)

C: A living room featuring a projection screen and recording equipment

Q_1: Is the picture in color?
A_1: Yes.

Q_2: Can you see any furniture?
A_2: Yes.

Q_3: Is there a couch in the room?
A_3: Yes.

Q_4: What color is the couch?
A_4: Blue.

A: Yes.
GT: Yes.

(c)

C: 2 guys work together to play a game of tennis.

Q_1: Is this picture in color?
A_1: Yes.

Q_2: Are they wearing hats?
A_2: Both are.

Q_3: What color is the court?
A_3: Blue.

A: Just hats.
GT: Just hats.

(d)

C: A little kid is wearing a part of skies.

Q_1: Is the photo in color?
A_1: Yes.

Q_2: Is there only one kid?
A_2: Yes.

A: Yes.
GT: Yes.

Fig. 9. Visualization results of iterative context-aware graph inference. It shows the word-level attention on question Q, and dynamic graph inference of the top-2 attended objects (red and blue bounding boxes) in image I. The number on each edge denotes the connection weight, displaying the message influence propagated from neighbors. There are some abbreviations as follows: question (Q), generated answer (A), caption (C) and the ground-truth (GT).

model RAA-Net [26]; hence, we experimented with RAA-Net and our graph model. We also tested RAA-Net-Distill with our distillation scheme.

As shown in Table 7, compared with either CAG or RAA-Net, the ensemble model **CAG+RAA-Net** shows an obvious performance improvement. For example, compared with CAG, the MRR value of "CAG+RAA-Net" increases from 64.57% to 67.75% and R@1 from 51.12% to 55.66%. After distillation, the performance of the ensemble model **(CAG + RAA-Net)-Distll** further improves with the best Mean value of 3.68, MRR value of 68.35, and R@1 value of 56.24. **ReDAN - 4 Ens** [25] integrated the probability scores of four ReDAN models. Our model only integrates two models: one graph-based model (CAG) and one attention model (RAA-Net). Except for NDCG, our ensemble model outperforms **ReDAN - 4 Ens** [25] on other metrics. The experimental results demonstrated that the ensemble model effectively improves the performance. It is feasible to explore more ensemble methods with graph and attention-based models.

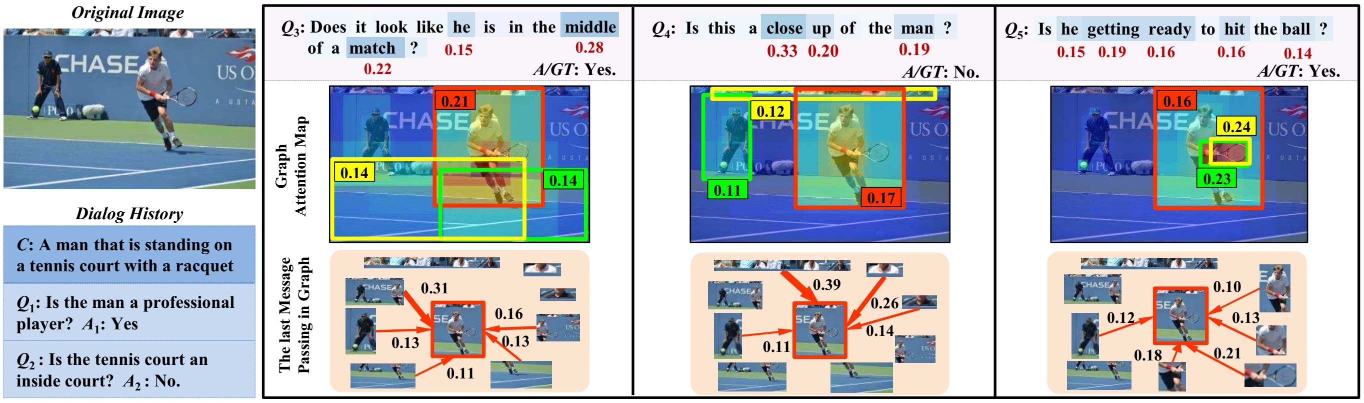

Original Image
Dialog History
C: A man that is standing on a tennis court with a racquet
Q1: Is the man a professional player? A1: Yes
Q2: Is the tennis court an inside court? A2: No.
Q3: Does it look like he is in the middle of a match ?
0.15 0.28 0.22
A/GT: Yes.
Q4: Is this a close up of the man ?
0.33 0.20 0.19
A/GT: No.
Q5: Is he getting ready to hit the ball ?
0.15 0.19 0.16 0.16 0.14
A/GT: Yes.
Graph Attention Map
The last Message Passing in Graph
0.21 0.14 0.14
0.12 0.11 0.17
0.16 0.24 0.23
0.31 0.16 0.13 0.13 0.11
0.39 0.26 0.14 0.11
0.10 0.12 0.13 0.18 0.21

Fig. 10. Visualization result of a progressive multi-round dialog inference. Each column shows a graph attention map and the last step of message passing process of a salient object. In these graph attention maps, bounding boxes correspond to the top-3 attended object nodes in the final graph, and the numbers along with the bounding boxes represent the node attention weights.

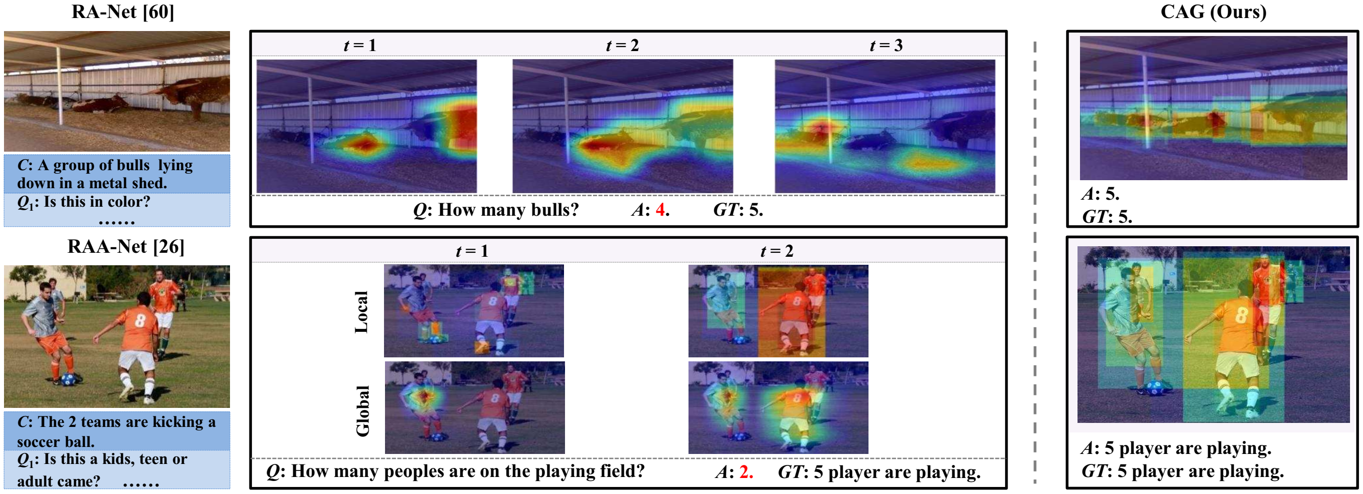

RA-Net [60]
C: A group of bulls lying down in a metal shed.
Q1: Is this in color?
……
t = 1
t = 2
t = 3
Q: How many bulls? A: 4. GT: 5.
CAG (Ours)
A: 5.
GT: 5.
RAA-Net [26]
C: The 2 teams are kicking a soccer ball.
Q1: Is this a kids, teen or adult came? ……
Local
Global
Q: How many peoples are on the playing field? A: 2. GT: 5 player are playing.
A: 5 player are playing.
GT: 5 player are playing.

Fig. 11. Visualization examples of attentive objects of images in VisDial v0.9. RA-Net [60] updates the global attention map at each step $t \in [1, T]$. RAA-Net [26] considers both global and local (object-level) attention maps. CAG only calculates the visual attention map at the object-level after T times graph inference. In our attention maps, we display the attentive regions merged with the bounding boxes of the weighted objects.

TABLE 7
Ensemble results of the graph-based and attention-based model on VisDial v1.0 val.

Types	Model	Mean↓	NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑
Attention	ReDAN [25]	4.05	59.32	64.21	50.60	81.39	90.26
	RAA-Net [26]	4.18	56.87	64.12	50.65	80.92	89.73
	RAA-Net-Distill	4.03	57.92	64.93	51.41	81.92	90.41
Graph	CAG	4.01	58.09	64.57	51.12	81.43	90.30
	CAG-Distill	3.94	59.26	65.38	51.94	82.14	90.75
Ensemble	ReDAN‡- 4 Ens [25]	3.82	60.53	65.30	51.67	82.40	91.09
	CAG+RAA	3.75	58.76	67.75	55.66	82.68	91.18
	CAG+RAA-Distill	3.70	59.07	68.06	56.01	83.01	91.39
	CAG-Distill+RAA	3.72	59.20	68.01	55.93	83.04	91.37
	(CAG+RAA)-Distill	**3.68**	**59.91**	**68.35**	**56.25**	**83.54**	**91.62**

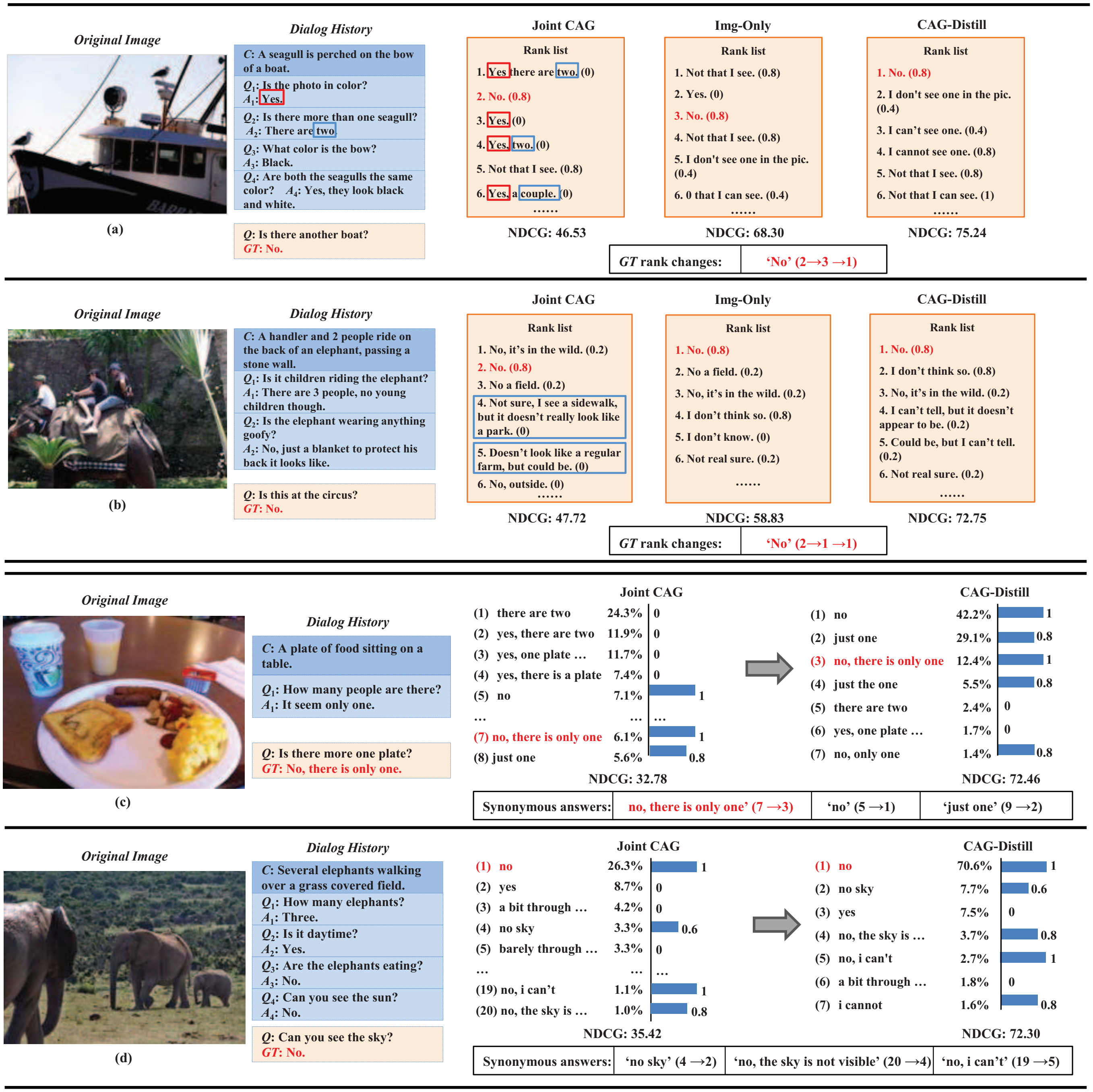
5.5 Qualitative Results

5.5.1 *Iterative Graph Inference*

To demonstrate the interpretability of our solution, we display an iterative graph inference example in Fig. 9. Two most salient objects ("snowboarder" and "pants") in the graph attention maps are selected to display the inference processes. At iteration step $t = 1$, the question focuses on the word "snowboarder" (target). By reviewing the dialog context, "snowboarder" is related with "midair" and "photographer". The above two salient objects receive messages from their relevant neighbor object nodes. Then at iteration step $t = 2$, the question changes the attention to both words of "snowboarder" and "wearing" (related objects). These two object nodes dynamically update their neighbor nodes under the guidance of the current question command $q_w^{(t=2)}$. At the last step $t = 3$, question Q focuses on the word "wearing" (relation). The edge connections in the graph are further refined by receiving messages from wearing-related nodes. Through multi-step message passing, our context-aware graph progressively finds out much more implicit question-related visual and textual semantics. Finally, the global image graph attention map also demonstrates the effectiveness of the graph inference.

5.5.2 *Relational Reasoning in Multi-round QA pairs*

The example in Fig. 10 describes relational reasoning in multi-round QA pairs. The edge relationships and the nodes' attention weights dynamically vary corresponding to the current question. Our context-aware graph effectively models these dynamical relations via the adaptive top-K



(a)

C: A seagull is perched on the bow of a boat.
Q_1: Is the photo in color? A_1: Yes.
Q_2: Is there more than one seagull? A_2: There are two
Q_3: What color is the bow? A_3: Black.
Q_4: Are both the seagulls the same color? A_4: Yes, they look black and white.

Q: Is there another boat?
GT: No.

Joint CAG — Rank list
1. Yes there are two. (0)
2. No. (0.8)
3. Yes. (0)
4. Yes, two. (0)
5. Not that I see. (0.8)
6. Yes, a couple. (0)
……
NDCG: 46.53

Img-Only — Rank list
1. Not that I see. (0.8)
2. Yes. (0)
3. No. (0.8)
4. Not that I see. (0.8)
5. I don't see one in the pic. (0.4)
6. 0 that I can see. (0.4)
……
NDCG: 68.30

CAG-Distill — Rank list
1. No. (0.8)
2. I don't see one in the pic. (0.4)
3. I can't see one. (0.4)
4. I cannot see one. (0.8)
5. Not that I see. (0.8)
6. Not that I can see. (1)
……
NDCG: 75.24

GT rank changes:	'No' (2→3 →1)

(b)

C: A handler and 2 people ride on the back of an elephant, passing a stone wall.
Q_1: Is it children riding the elephant? A_1: There are 3 people, no young children though.
Q_2: Is the elephant wearing anything goofy? A_2: No, just a blanket to protect his back it looks like.

Q: Is this at the circus?
GT: No.

Joint CAG — Rank list
1. No, it's in the wild. (0.2)
2. No. (0.8)
3. No a field. (0.2)
4. Not sure, I see a sidewalk, but it doesn't really look like a park. (0)
5. Doesn't look like a regular farm, but could be. (0)
6. No, outside. (0)
……
NDCG: 47.72

Img-Only — Rank list
1. No. (0.8)
2. No a field. (0.2)
3. No, it's in the wild. (0.2)
4. I don't think so. (0.8)
5. I don't know. (0)
6. Not real sure. (0.2)
……
NDCG: 58.83

CAG-Distill — Rank list
1. No. (0.8)
2. I don't think so. (0.8)
3. No, it's in the wild. (0.2)
4. I can't tell, but it doesn't appear to be. (0.2)
5. Could be, but I can't tell. (0.2)
6. Not real sure. (0.2)
……
NDCG: 72.75

GT rank changes:	'No' (2→1 →1)

(c)

C: A plate of food sitting on a table.
Q_1: How many people are there? A_1: It seem only one.

Q: Is there more one plate?
GT: No, there is only one.

Joint CAG

(1) there are two	24.3%	0
(2) yes, there are two	11.9%	0
(3) yes, one plate …	11.7%	0
(4) yes, there is a plate	7.4%	0
(5) no	7.1%	1
…	…	…
(7) no, there is only one	6.1%	1
(8) just one	5.6%	0.8

NDCG: 32.78

CAG-Distill

(1) no	42.2%	1
(2) just one	29.1%	0.8
(3) no, there is only one	12.4%	1
(4) just the one	5.5%	0.8
(5) there are two	2.4%	0
(6) yes, one plate …	1.7%	0
(7) no, only one	1.4%	0.8

NDCG: 72.46

Synonymous answers:	no, there is only one' (7 →3)	'no' (5 →1)	'just one' (9 →2)

(d)

C: Several elephants walking over a grass covered field.
Q_1: How many elephants? A_1: Three.
Q_2: Is it daytime? A_2: Yes.
Q_3: Are the elephants eating? A_3: No.
Q_4: Can you see the sun? A_4: No.

Q: Can you see the sky?
GT: No.

Joint CAG

(1) no	26.3%	1
(2) yes	8.7%	0
(3) a bit through …	4.2%	0
(4) no sky	3.3%	0.6
(5) barely through …	3.3%	0
…	…	…
(19) no, i can't	1.1%	1
(20) no, the sky is …	1.0%	0.8

NDCG: 35.42

CAG-Distill

(1) no	70.6%	1
(2) no sky	7.7%	0.6
(3) yes	7.5%	0
(4) no, the sky is …	3.7%	0.8
(5) no, i can't	2.7%	1
(6) a bit through …	1.8%	0
(7) i cannot	1.6%	0.8

NDCG: 72.30

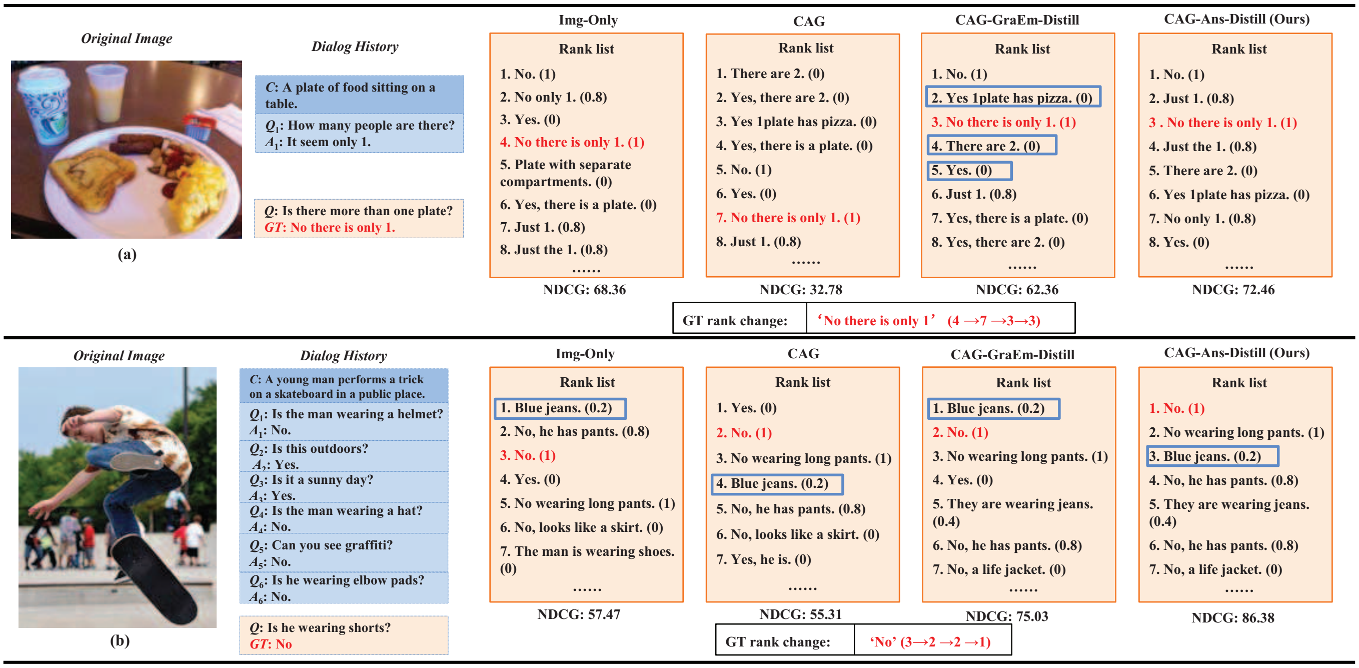
Synonymous answers:	'no sky' (4 →2)	'no, the sky is not visible' (20 →4)	'no, i can't' (19 →5)

Fig. 12. Qualitative examples of predicted ranking lists. The format of answer rank list in (a) and (b) is fixed to "predicted rank order - answer - (annotated relevance score)", and "predicted rank order - answer - probability score - annotated relevance score" in (c) and (d). Ground-truth (*GT*) answers are marked with red fonts.

message passing module. Each node only receives strong messages from the most relevant nodes. The global image attention maps at different rounds further validate the adaptability of our graph on relational reasoning.

5.5.3 Attentive Objects Analysis

In this subsection, we display the visualization examples of attentive objects of images in VisDial v0.9. As shown in Fig. 11 (a), compared with global attention in RA-Net [60], local attention with fine-grained object-level visual clues takes a more positive effect. Our attention map performs remarkably better with exact attentive location, explicit boundary, and a highly responsive heat map of the noun (five "bulls") in the ground-truth answer. RAA-Net [26] is a typical work involving both global and local visual attentions. As shown in Fig. 11 (b), except for obscure and inaccurate attentive regions appearing in the global map, local attention at the object-level in RAA-Net [26] just focuses on two salient objects covering four people. There are actually five people in the image. Inspired by our graph inference at the object-level, after graph learning, the proposed CAG adaptively assigns important weights to multiple objects with bounding boxes covering five people.



C: A plate of food sitting on a table.
Q$_1$: How many people are there?
A$_1$: It seem only 1.

Q: Is there more than one plate?
GT: No there is only 1.

Img-Only — Rank list
1. No. (1)
2. No only 1. (0.8)
3. Yes. (0)
4. No there is only 1. (1)
5. Plate with separate compartments. (0)
6. Yes, there is a plate. (0)
7. Just 1. (0.8)
8. Just the 1. (0.8)
……
NDCG: 68.36

CAG — Rank list
1. There are 2. (0)
2. Yes, there are 2. (0)
3. Yes 1plate has pizza. (0)
4. Yes, there is a plate. (0)
5. No. (1)
6. Yes. (0)
7. No there is only 1. (1)
8. Just 1. (0.8)
……
NDCG: 32.78

CAG-GraEm-Distill — Rank list
1. No. (1)
2. Yes 1plate has pizza. (0)
3. No there is only 1. (1)
4. There are 2. (0)
5. Yes. (0)
6. Just 1. (0.8)
7. Yes, there is a plate. (0)
8. Yes, there are 2. (0)
……
NDCG: 62.36

CAG-Ans-Distill (Ours) — Rank list
1. No. (1)
2. Just 1. (0.8)
3 . No there is only 1. (1)
4. Just the 1. (0.8)
5. There are 2. (0)
6. Yes 1plate has pizza. (0)
7. No only 1. (0.8)
8. Yes. (0)
……
NDCG: 72.46

GT rank change: ‘No there is only 1’ (4 →7 →3→3)

C: A young man performs a trick on a skateboard in a public place.
Q$_1$: Is the man wearing a helmet?
A$_1$: No.
Q$_2$: Is this outdoors?
A$_2$: Yes.
Q$_3$: Is it a sunny day?
A$_3$: Yes.
Q$_4$: Is the man wearing a hat?
A$_4$: No.
Q$_5$: Can you see graffiti?
A$_5$: No.
Q$_6$: Is he wearing elbow pads?
A$_6$: No.

Q: Is he wearing shorts?
GT: No

Img-Only — Rank list
1. Blue jeans. (0.2)
2. No, he has pants. (0.8)
3. No. (1)
4. Yes. (0)
5. No wearing long pants. (1)
6. No, looks like a skirt. (0)
7. The man is wearing shoes. (0)
……
NDCG: 57.47

CAG — Rank list
1. Yes. (0)
2. No. (1)
3. No wearing long pants. (1)
4. Blue jeans. (0.2)
5. No, he has pants. (0.8)
6. No, looks like a skirt. (0)
7. Yes, he is. (0)
……
NDCG: 55.31

CAG-GraEm-Distill — Rank list
1. Blue jeans. (0.2)
2. No. (1)
3. No wearing long pants. (1)
4. Yes. (0)
5. They are wearing jeans. (0.4)
6. No, he has pants. (0.8)
7. No, a life jacket. (0)
……
NDCG: 75.03

CAG-Ans-Distill (Ours) — Rank list
1. No. (1)
2. No wearing long pants. (1)
3. Blue jeans. (0.2)
4. No, he has pants. (0.8)
5. They are wearing jeans. (0.4)
6. No, he has pants. (0.8)
7. No, a life jacket. (0)
……
NDCG: 86.38

GT rank change: ‘No’ (3→2 →2 →1)

Fig. 13. Visualization examples of different distillation schemes. The ground-truth answers (GT) are marked with red fonts. Compared with others, CAG-Ans-Distill prefers to rank closely correct answers at the top of score list. The candidate answers in blue boxes contain historical noises.

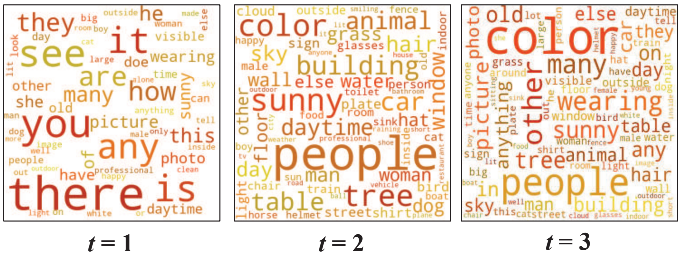


Fig. 14. Visualization of attentive word cloud of all the questions $\{Q\}$ at different iteration steps on VisDial v1.0.

In summary, the essential difference is that our work obtains the visual attention map through node attention at the object-level after graph inference. This can be explained in two aspects: (1) object-level nodes introduce explicit location and boundary, and (2) the early iterative graph inference is helpful to the latter weight assignment to crucial nodes.

5.5.4 Attentive Word Clouds Analysis

Here, we display the visualization of attentive word clouds on VisDial v1.0. Fig. 14 describes the word-level attention distribution of question Q at different iteration steps. At iteration step $t = 1$, the proposed CAG inclines to pronouns in the questions, *e.g.*, "there", "it", "you", "they". CAG tries to tackle the textual co-reference in the initial relational reasoning. Then, at step $t = 2$, CAG prefers to attend nouns related to target objects or associated objects in the image, *e.g.*, "people", "building", "tree". This means CAG trends to infer the relationships between different related objects, namely visual-reference. At the time step $t = 3$, the model considers the words that describe the attributes or relations of the objects, *e.g.*, "color", "wearing", "other", "on". All these appearances indicate that we reasonably and actively promote the iterative inference process using the context-aware graph CAG.

5.5.5 Effect of Knowledged Distillation: More Reasonable Answer Ranking

In this subsection, we first visualize four examples of the rank changes of candidate answers among Img-Only CAG, Joint CAG (*i.e.*. abbreviated CAG) and CAG-Distill models. As shown in Fig. 12 (a), compared with Joint CAG, CAG-Distill has a powerful correction capability with the knowledge learned from Img-Only model; in Fig. 12 (b), the Joint CAG which explicitly learns the historical clues, prefers to the long answers with rich semantics; finally, in Figs. 12 (c) and (d), dense blue status bars appear in the rank list of CAG-Ditill, it indicates that after knowledge distilling, the predicted answers with top rank orders are more relevant to the GT answers. Synonymous answers of the ground-truth can be ranked at top positions.

Secondly, we test both **CAG-GraEm-Distill** and **CAG-Ans-Distill** schemes. As shown in Fig. 13 (a), CAG introduces history noises (number "2" and phase "1/a plate") and fails to infer the exact answer (listing at rank 7). **CAG-GraEm-Distill** improves it, where the exact answer appears at rank-1 but there are oppositely wrong answers at ranks 2, 4, and 5. In contrast, **CAG-Ans-Distill** further improves the answer score list, in which the top four answers are closely related to the correct answer. As the example shown in Fig. 13 (b), history is not needed. It is theoretically possible to obtain the correct answer through the Img-Only model. However, the Img-only model fails and prefers a rich semantic phase to describe the pants ("Blue jeans") rather than "no shorts". Influenced by the teacher model (Img-Only), both **CAG** and **CAG-GraEm-Distill** fail. **CAG-Ans-Distill** improves it, distilling the knowledge at the answer probability level.

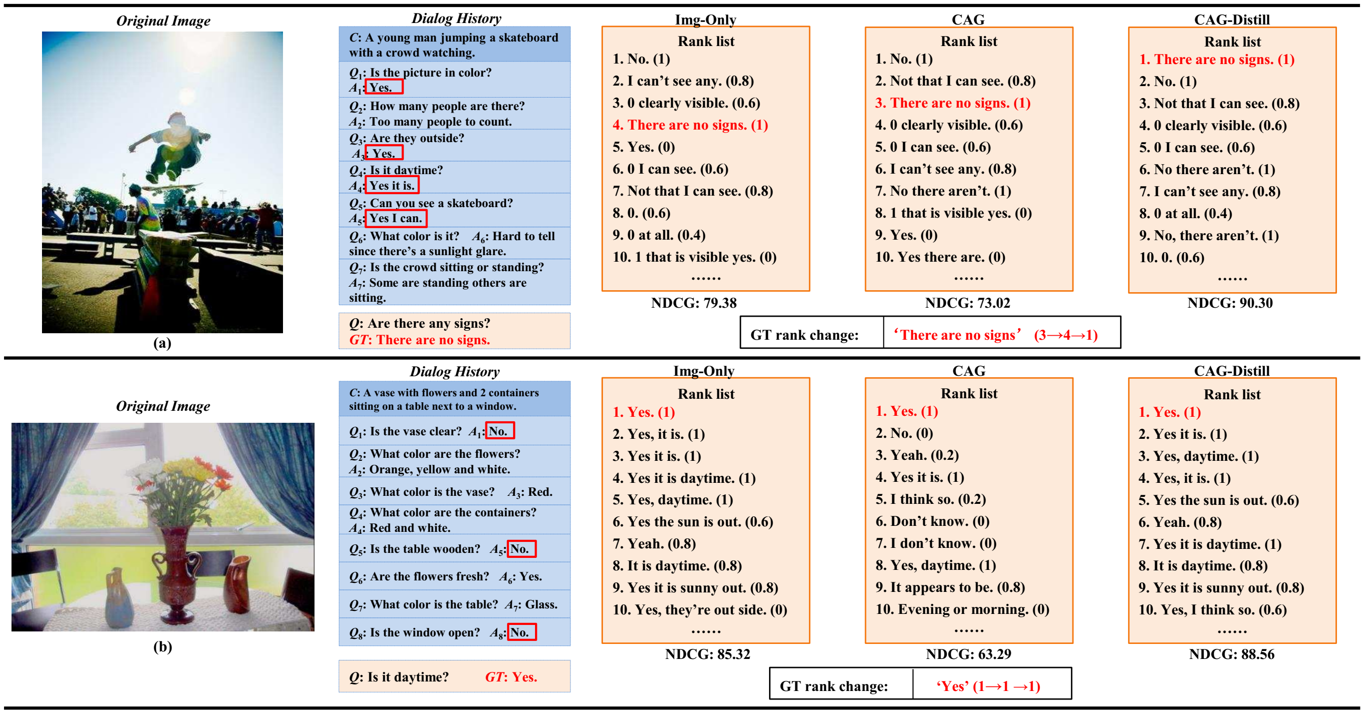

Original Image
(a)
Dialog History
C: A young man jumping a skateboard with a crowd watching.
Q_1: Is the picture in color?
A_1: Yes.
Q_2: How many people are there?
A_2: Too many people to count.
Q_3: Are they outside?
A_3: Yes.
Q_4: Is it daytime?
A_4: Yes it is.
Q_5: Can you see a skateboard?
A_5: Yes I can.
Q_6: What color is it? A_6: Hard to tell since there's a sunlight glare.
Q_7: Is the crowd sitting or standing?
A_7: Some are standing others are sitting.
Q: Are there any signs?
GT: There are no signs.
Img-Only
Rank list
1. No. (1)
2. I can't see any. (0.8)
3. 0 clearly visible. (0.6)
4. There are no signs. (1)
5. Yes. (0)
6. 0 I can see. (0.6)
7. Not that I can see. (0.8)
8. 0. (0.6)
9. 0 at all. (0.4)
10. 1 that is visible yes. (0)
……
NDCG: 79.38
CAG
Rank list
1. No. (1)
2. Not that I can see. (0.8)
3. There are no signs. (1)
4. 0 clearly visible. (0.6)
5. 0 I can see. (0.6)
6. I can't see any. (0.8)
7. No there aren't. (1)
8. 1 that is visible yes. (0)
9. Yes. (0)
10. Yes there are. (0)
……
NDCG: 73.02
CAG-Distill
Rank list
1. There are no signs. (1)
2. No. (1)
3. Not that I can see. (0.8)
4. 0 clearly visible. (0.6)
5. 0 I can see. (0.6)
6. No there aren't. (1)
7. I can't see any. (0.8)
8. 0 at all. (0.4)
9. No, there aren't. (1)
10. 0. (0.6)
……
NDCG: 90.30
GT rank change:
'There are no signs' (3→4→1)
Original Image
(b)
Dialog History
C: A vase with flowers and 2 containers sitting on a table next to a window.
Q_1: Is the vase clear? A_1: No.
Q_2: What color are the flowers?
A_2: Orange, yellow and white.
Q_3: What color is the vase? A_3: Red.
Q_4: What color are the containers?
A_4: Red and white.
Q_5: Is the table wooden? A_5: No.
Q_6: Are the flowers fresh? A_6: Yes.
Q_7: What color is the table? A_7: Glass.
Q_8: Is the window open? A_8: No.
Q: Is it daytime? GT: Yes.
Img-Only
Rank list
1. Yes. (1)
2. Yes, it is. (1)
3. Yes it is. (1)
4. Yes it is daytime. (1)
5. Yes, daytime. (1)
6. Yes the sun is out. (0.6)
7. Yeah. (0.8)
8. It is daytime. (0.8)
9. Yes it is sunny out. (0.8)
10. Yes, they're out side. (0)
……
NDCG: 85.32
CAG
Rank list
1. Yes. (1)
2. No. (0)
3. Yeah. (0.2)
4. Yes it is. (1)
5. I think so. (0.2)
6. Don't know. (0)
7. I don't know. (0)
8. Yes, daytime. (1)
9. It appears to be. (0.8)
10. Evening or morning. (0)
……
NDCG: 63.29
CAG-Distill
Rank list
1. Yes. (1)
2. Yes it is. (1)
3. Yes, daytime. (1)
4. Yes, it is. (1)
5. Yes the sun is out. (0.6)
6. Yeah. (0.8)
7. Yes it is daytime. (1)
8. It is daytime. (0.8)
9. Yes it is sunny out. (0.8)
10. Yes, I think so. (0.6)
……
NDCG: 88.56
GT rank change:
'Yes' (1→1 →1)

Fig. 15. Positive examples of our CAG models inhibiting the negative influences of high-frequency words in history.

At last, Fig. 15 visualizes some positive examples inhibiting the negative influences of high-frequency words in history. These results further validate the generalization and discrimination abilities of the CAG model. CAG-Distill improves the answer preciseness and prefers relevant answers with high ranks and confident probability scores.

6 Conclusion

In this paper, we propose a fine-grained Context-Aware Graph (CAG) neural network for visual dialog, which contains both visual-objects and textual-history context semantics. An adaptive top-K message passing mechanism is proposed to iteratively explore the context-aware representations of nodes and update the edge relationships for a better answer inferring. Our solution is a dynamic directed-graph inference process. Furthermore, to refine both good generalization ability and discrimination ability of the model, the proposed visual-aware knowledge distillation on the graph learning takes effect. CAG-Distill has excellent robustness. Experimental results on the VisDial v0.9 and v1.0 datasets validate the effectiveness of the proposed approach, and display explainable visualization results.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61725203, Grant 62020106007, Grant 61876058, and Fundamental Research Funds for the Central Universities under Grant JZ2020HGTB0020.

References

[1] F. Zhao, J. Li, J. Zhao, and J. Feng, "Weakly supervised phrase localization with multi-scale anchored transformer network," in *Proc. IEEE Conf. Comp. Vis. Pattern Recog.*, 2018, pp. 5696–5705.

[2] Y. Niu, H. Zhang, Z. Lu, and S. Chang, "Variational context: Exploiting visual and textual context for grounding referring expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2019.

[3] S. Yang, G. Li, and Y. Yu, "Relationship-embedded representation learning for grounding referring expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2020.

[4] Z. Chen, P. Wang, L. Ma, K. K. Wong, and Q. Wu, "Cops-ref: A new dataset and task on compositional referring expression comprehension," in *Proc. IEEE Conf. Comp. Vis. Pattern Recog.*, 2020, pp. 10 083–10 092.

[5] Z. Zhang, Z. Zhao, Y. Zhao, Q. Wang, H. Liu, and L. Gao, "Where does it exist: Spatio-temporal video grounding for multi-form sentences," in *Proc. IEEE Conf. Comp. Vis. Pattern Recog.*, 2020, pp. 10 665–10 674.

[6] L. Huang, W. Wang, J. Chen, and X. Wei, "Attention on attention for image captioning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 4633–4642.

[7] Z.-J. Zha, D. Liu, H. Zhang, Y. Zhang, and F. Wu, "Context-aware visual policy network for fine-grained image captioning," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2019.

[8] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proc. IEEE Conf. Comp. Vis. Pattern Recog.*, 2020, pp. 10 968–10 977.

[9] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proc. IEEE Conf. Comp. Vis. Pattern Recog.*, 2020, pp. 10 575–10 584.

[10] D. Guo, Y. Wang, P. Song, and M. Wang, "Recurrent relational memory network for unsupervised image captioning," in *Proc. 29th Int. J. Conf. Artif. Intell.*, 2020, pp. 920–926.

[11] J. Zhang and Y. Peng, "Object-aware aggregation with bidirectional temporal graph for video captioning," in *Proc. IEEE Conf. Comp. Vis. Pattern Recog.*, 2019, pp. 8327–8336.

[12] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2425–2433.

[13] Q. Wu, C. Shen, P. Wang, A. Dick, and A. v. d. Hengel, "Image captioning and visual question answering based on attributes and external knowledge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1367–1381, 2018.

[14] J. Liang, L. Jiang, L. Cao, Y. Kalantidis, L. Li, and A. G. Hauptmann, "Focal visual-text attention for memex question answering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1893–1908, 2019.

[15] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim, "TGIF-QA: toward spatio-temporal reasoning in visual question answering," in *Proc. IEEE Conf. Comp. Vis. Pattern Recog.*, 2017, pp. 1359–1367.

[16] J. Lei, L. Yu, T. L. Berg, and M. Bansal, "TVQA+: spatio-temporal grounding for video question answering," in *Proc. Conf. Assoc. Comput. Linguistics*, 2020, pp. 8211–8225.

[17] J. Kim, M. Ma, T. X. Pham, K. Kim, and C. D. Yoo, "Modality shifting attention network for multi-modal video question answering," in *Proc. IEEE Conf. Comp. Vis. Pattern Recog.*, 2020, pp. 10 103–10 112.

[18] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proc. IEEE Conf. Comp. Vis. Pattern Recog.*, 2016.

[19] H. Fan and Y. Yang, "Person tube retrieval via language description," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 10 754–10 761.

[20] H. de Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. Courville, "Guesswhat?! visual object discovery through multi-modal dialogue," in *Proc. IEEE Conf. Comp. Vis. Pattern Recog.*, 2017, pp. 5503–5512.

[21] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. F. Moura, D. Parikh, and D. Batra, "Visual dialog," in *Proc. IEEE Conf. Comp. Vis. Pattern Recog.*, 2017, pp. 326–335.

[22] S. Kottur, J. M. F. Moura, D. Parikh, D. Batra, and M. Rohrbach, "Clevr-dialog: A diagnostic dataset for multi-round reasoning in visual dialog," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 582–595.

[23] Z. Zheng, W. Wang, S. Qi, and S.-C. Zhu, "Reasoning visual dialogs with structural and partial observations," in *Proc. IEEE Conf. Comp. Vis. Pattern Recog.*, 2019, pp. 6669–6678.

[24] I. Schwartz, S. Yu, T. Hazan, and A. G. Schwing, "Factor graph attention," in *Proc. IEEE Conf. Comp. Vis. Pattern Recog.*, 2019, pp. 2039–2048.

[25] Zhe Gan and Yu Cheng and Ahmed El Kholy and Linjie Li and Jingjing Liu and Jianfeng Gao, "Multi-step Reasoning via Recurrent Dual Attention for Visual Dialog," in *Proc. Conf. Assoc. Comput. Linguistics*, 2019, pp. 6463–6474.

[26] D. Guo, H. Wang, S. Wang, and M. Wang, "Textual-visual reference-aware attention network for visual dialog," *IEEE Trans. Image Process.*, vol. 29, pp. 6655–6666, 2020.

[27] Shubham Agarwal and Trung Bui and Joon-Young Lee and Ioannis Konstas and Verena Rieser, "History for Visual Dialog: Do we really need it?" in *Proc. Conf. Assoc. Comput. Linguistics*, 2020, pp. 8182–8197.

[28] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1839–1848.

[29] X. Jiang, J. Yu, Z. Qin, Y. Zhuang, X. Zhang, Y. Hu, and Q. Wu, "Dualvd: An adaptive dual encoding model for deep visual understanding in visual dialogue," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 11 125–11 132.

[30] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015.

[31] B. Pan, H. Cai, D. Huang, K. Lee, A. Gaidon, E. Adeli, and J. C. Niebles, "Spatio-temporal graph for video captioning with knowledge distillation," in *Proc. IEEE Conf. Comp. Vis. Pattern Recog.*, 2020, pp. 10 867–10 876.

[32] D. Guo, H. Wang, H. Zhang, Z. Zha, and M. Wang, "Iterative context-aware graph inference for visual dialog," in *Proc. IEEE Conf. Comp. Vis. Pattern Recog.*, 2020, pp. 10 052–10 061.

[33] J. Lu, A. Kannan, J. Yang, D. Parikh, and D. Batra, "Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 314–324.

[34] Q. Wu, P. Wang, C. Shen, I. Reid, and A. van den Hengel, "Are you talking to me? reasoned visual dialog generation through adversarial learning," in *Proc. IEEE Conf. Comp. Vis. Pattern Recog.*, 2018, pp. 6106–6115.

[35] D. Guo, H. Wang, and M. Wang, "Dual visual attention network for visual dialog," in *Proc. 28th Int. J. Conf. Artif. Intell.*, 2019, pp. 4989–4995.

[36] P. H. Seo, A. Lehrmann, B. Han, and L. Sigal, "Visual reference resolution using attention memory for visual dialog," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 3719–3729.

[37] S. Kottur, J. M. F. Moura, D. Parikh, D. Batra, and M. Rohrbach, "Visual coreference resolution in visual dialog using neural module networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 153–169.

[38] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Neural module networks," in *Proc. IEEE Conf. Comp. Vis. Pattern Recog.*, 2016, pp. 39–48.

[39] Y. Niu, H. Zhang, M. Zhang, J. Zhang, Z. Lu, and J.-R. Wen, "Recursive visual attention in visual dialog," in *Proc. IEEE Conf. Comp. Vis. Pattern Recog.*, 2019, pp. 6679–6688.

[40] P. Wang, Q. Wu, J. Cao, C. Shen, L. Gao, and A. v. d. Hengel, "Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks," in *Proc. IEEE Conf. Comp. Vis. Pattern Recog.*, 2019, pp. 1960–1968.

[41] Y. Liu, R. Wang, S. Shan, and X. Chen, "Structure inference net: Object detection using scene-level context and instance-level relationships," in *Proc. IEEE Conf. Comp. Vis. Pattern Recog.*, 2018, pp. 6985–6994.

[42] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE Conf. Comp. Vis. Pattern Recog.*, 2019, pp. 3595–3603.

[43] J. Gu, H. Zhao, Z. Lin, S. Li, J. Cai, and M. Ling, "Scene graph generation with external knowledge and image reconstruction," in *Proc. IEEE Conf. Comp. Vis. Pattern Recog.*, 2019, pp. 1969–1978.

[44] Y. Peng and J. Chi, "Unsupervised cross-media retrieval using domain adaptation with scene graph," *IEEE Trans. Circuits Syst. Video Technol.*, pp. 1–1, 2019.

[45] D. Teney, L. Liu, and A. van den Hengel, "Graph-structured representations for visual question answering," in *Proc. IEEE Conf. Comp. Vis. Pattern Recog.*, 2017, pp. 3233–3241.

[46] W. Norcliffe-Brown, S. Vafeias, and S. Parisot, "Learning conditioned graph structures for interpretable visual question answering," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 8344–8353.

[47] J. Zhang, Y. Kalantidis, M. Rohrbach, M. Paluri, A. Elgammal, and M. Elhoseiny, "Large-scale visual relationship understanding," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2019, pp. 9185–9194.

[48] C. Bucila, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2006, pp. 535–541.

[49] D. Lopez-Paz, L. Bottou, B. Schölkopf, and V. Vapnik, "Unifying distillation and privileged information," in *Proc. Int. Conf. Learn. Representations*, 2016.

[50] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural Networks*, vol. 22, no. 5-6, pp. 544–557, 2009.

[51] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in *Proc. IEEE Conf. Comp. Vis. Pattern Recog.*, 2016, pp. 2827–2836.

[52] Yu Wu and Lu Jiang and Yi Yang, "Revisiting EmbodiedQA: A Simple Baseline and Beyond," *IEEE Trans. Image Process.*, vol. 29, pp. 3984–3992, 2020.

[53] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comp. Vis. Pattern Recog.*, 2018, pp. 6077–6086.

[54] D. Guo, C. Xu, and D. Tao, "Image-question-answer synergistic network for visual dialog," in *Proc. IEEE Conf. Comp. Vis. Pattern Recog.*, 2019, pp. 10 434–10 443.

[55] G. Kang, J. Lim, and B. Zhang, "Dual attention networks for visual reference resolution in visual dialog," pp. 2024–2033, 2019.

[56] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[57] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Meth. Nat. Language Process.*, 2014, pp. 1532–1543.

[58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.

[59] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[60] Hehe Fan and Linchao Zhu and Yi Yang and Fei Wu, "Recurrent Attention Network with Reinforced Generator for Visual Dialog," *ACM Trans. Multim. Comput. Commun. Appl.*, vol. 16, no. 3, pp. 78:1–78:16, 2020.

[61] T. Yang, Z.-J. Zha, and H. Zhang, "Making history matter: History-advantage sequence training for visual dialog," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 2561–2569.

[62] Y. Wang, L. Wang, Y. Li, D. He, and T. Liu, "A theoretical analysis of NDCG type ranking measures," in *COLT*, 2013, pp. 25–54.

Dan Guo received the B.E. degree in computer science and technology from Yangtze University, China, in 2004, and the Ph.D. degree in system analysis and integration from the Huazhong University of Science and Technology, China, in 2010. She is currently an Professor at the School of Computer and Information, Hefei University of Technology. Her research interests include computer vision, machine learning, and intelligent multimedia content analysis.

Hui Wang received the B.E. degree in computer science and technology from Hefei University of Technology, China, in 2018. He is currently pursuing the Ph.D. degree in computer technology with Hefei University of Technology, China. His current research interests include computer vision and deep learning.

Meng Wang (M09-SM17-F21) is a professor at Hefei University of Technology, China. He received his B.E. degree and Ph.D. degree in the Special Class for the Gifted Young and the Department of Electronic Engineering and Information Science from the University of Science and Technology of China (USTC), Hefei, China, in 2003 and 2008, respectively. His current research interests include multimedia content analysis, computer vision, and pattern recognition. He has authored over 200 book chapters, journal and conference papers in these areas. He is the recipient of the ACM SIGMM Rising Star Award 2014. He is an associate editor of IEEE Transactions on Knowledge and Data Engineering (IEEE TKDE), IEEE Transactions on Circuits and Systems for Video Technology (IEEE TCSVT), IEEE Transactions on Multimedia (IEEE TMM), and IEEE Transactions on Neural Networks and Learning Systems (IEEE TNNLS).