A. Supplementary Material

This supplementary document is organized as follows:

- We explain two different textual attention mechanisms in Sec. A.1, especially for the word-level attention on question Q at different iterative processes.
- Sec. A.2 elaborates the motivation of the fine-grained graph construction and the relational measurement between different nodes.
- In the paper, Figs. 5∼7 illustrate the visualization results on image I and question Q. Here, Sec. A.3 mainly demonstrates the influence of history-related context u in Fig. 9, and supplements additional qualitative results of visual reference in Fig. 10.

A.1. Textual Attention Mechanisms

- **Sentence-level attention on history** H: The latent attention variable z_h in Eq. 1 tackles the textual co-reference between q_s and U^H, where both q_s and U^H are sentence-level semantics.
- **Word-level attention on question** Q: For computing word-level attention on Q, the latent attention variable $z_q^{(t)}$ in Eq. 3 measures the question itself. We unfold the MLP operation in Eq. 3 as follows:

$$\begin{cases} f_q^{(t)} = tanh(W_{f_1}U^Q) \odot \sigma(W_{f_2}U^Q); \\ z_q^{(t)} = L2Norm(f_q^{(t)}(U^Q)), \end{cases} \tag{A1}$$

where $W_{f_1}, W_{f_2} \in \mathbb{R}^{d\times d}$ are learnable parameters.

The operation $f_q^{(t)}$ uses the tangent & sigmoid activation gates to learn a new word-level feature sequence of question Q. Then, the $L2Norm$ operation normalizes each word's new feature embedding vector on the feature dimension. With the L2 normalization, $z_q^{(t)} \in \mathbb{R}^{d\times m}$ can equitably evaluate each word in the word sequence of Q. Figs. 5∼7 (especially Fig. 7) validate the adaptability of the word-level attention in different iterative steps.

A.2. Fine-grained Graph Construction

A.2.1 Node components: visual and textual contexts

One argue maybe that why we realize the graph initialization with history-related context u and implement the joint visual-textual context learning? The motivation is that without history-related context u, the dialog agent can't understand the previous dialogue topic well, nor it can further solve the current visual reference well.

Technically, one challenge of visual dialog is to explore the latent relations among **image**, **history** and **question**. We reformulate the idea of the dynamic graph learning in the paper as follows. In iterative step t, as the definition of node $\mathcal{N}_i^{(t)} = [v_i; c_i^{(t)}]$, v_i denotes the visual feature of object obj_i, and $c_i^{(t)}$ records the relevant context related to obj_i. c_i considers both $\{v_i\}$ and u, and is guided by the question command $q_w^{(t)}$.

$$\begin{cases} \mathcal{N}_i^{(1)} = [v_i; c_i^{(1)}] = [\boldsymbol{v_i}; \boldsymbol{u}]; \\ c_i^{(t+1)} = \boldsymbol{MP}\Big(\boldsymbol{q_w^{(t)}} \odot \{\mathcal{N}_j^{(t)}\}|_{top-K} \succ \mathcal{N}_i^{(t)}\Big) \bowtie c_i^{(t)}; \\ \mathcal{N}_i^{(t+1)} = [v_i; c_i^{(t+1)}], \quad i \in [1, n], t \in [1, T], \end{cases} \tag{A2}$$

where $\mathcal{N}_i^{(1)}$ denotes the graph initialization, ***MP*** denotes the message passing calculation by Eq. 6, $\succ$ denotes the adjacent correlation matrix learning by Eq. 4, and $\bowtie$ means that the context $c_i^{(t)}$ is updated by Eq. 7.

The joint context learning of $c_i^{(t)}$ involving u plays an important role in the graph inference. Both ablation studies in Table 1 and qualitative results in Sec. A.3 detailed below demonstrate the effectiveness. In addition, Fig. 7 also verifies the significance of u in step t=1. The introduce of u is helpful to tackle the visual-textual co-reference related to the question, such as the parsing pronouns in the question ($e.g.$, "he", "it" and "there") and grounding the relevant objects in the image.

A.2.2 Adjacent correlation matrix learning

Another argue maybe that why impose the question command $q_w^{(t)}$ on only one node side of the matrix $A^{(t)}$ in Eq. 4, which is not a symmetrical operation as mutual correlation calculation. We define a classical mutual (symmetrical) correlation calculation as **CAG-DualQ** as follows:

$$\begin{cases} \textbf{CAG}: \\ A^{(t)} = (W_1\mathcal{N}^{(t)})^\top \Big[(W_2\mathcal{N}^{(t)}) \odot (W_3 q_w^{(t)})\Big]; \\ \textbf{CAG-DualQ}: \\ A^{(t)} = \Big[(W_1\mathcal{N}^{(t)}) \odot (W_3' q_w^{(t)})\Big]^\top \Big[(W_2\mathcal{N}^{(t)}) \odot (W_3 q_w^{(t)})\Big]. \end{cases} \tag{A3}$$

where W_1 and $W_2 \in \mathbb{R}^{d\times 2d}$, W_3 and $W_3' \in \mathbb{R}^{d\times d_w}$ are learnable parameters.

We implement the ablation study. As shown in Table 4, **CAG-DualQ** performs worse than **CAG**. It is interpretable. As illustrated in Fig. 8, the Y-axis of the matrix $A^{(t)}$ marks the receiving nodes, and the X-axis denotes the distributing nodes. To infer an exact answer, for a node, we use the question command $q_w^{(t)}$ to activate its neighbors. In other words, the i-th row of the matrix $A_i^{(t)}$ calculates the correlation weights of node $\mathcal{N}_i^{(t)}$ and its neighbors $\{\mathcal{N}_j^{(t)}\}$ under the only once guidance of $q_w^{(t)}$. It is reasonable to introduce the question cue on one node side of $A^{(t)}$.

Figure 9. Qualitative results of **CAG** and **CAG w/o *u***. Observing the graph attention map overlaying each image, the bounding boxes with the attention scores correspond to the top-3 relevant object nodes in the final graph. We pick out the top-2 objects to display the dynamic graph inference. **CAG** and **CAG w/o *u*** can refer to different top-2 objects. Without history context u, the agent could misunderstand the dialogue topic, and the visual reference cannot be solved well.

Model	Mean↓	MRR↑	R@1↑	R@5↑	R@10↑
CAG-DualQ	3.79	67.19	54.16	83.44	91.32
CAG	**3.75**	**67.56**	**54.64**	**83.72**	**91.48**

Table 4. Ablation studies of different adjacent correlation matrix learning strategies on VisDial val v0.9.

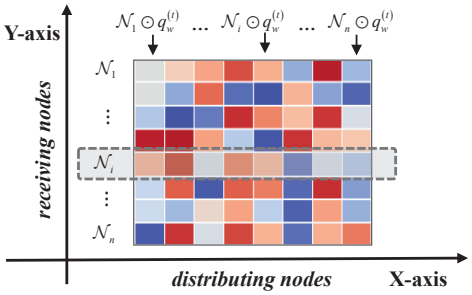

Figure 8. A schematic diagram of adjacent correlation matrix learning.

A.3. Additional Qualitative Results

A.3.1 Qualitative results of CAG vs. CAG w/o *u*

As the ablation study shown in Table 1, the performance of **CAG w/o *u*** drops a lot compared to **CAG**. Here, we provide explainable qualitative results in Fig. 9 to further validate the effectiveness of history-related context u in **CAG**. There are two different examples. As shown in Fig. 9 (a), for question Q: "He wearing helmet?", **CAG w/o *u*** directly locates words "he" (two people) and "helmet" in the image, and then the object "helmet" infers to a wrong "he" (the man who wears the helmet), while **CAG** consistently attends on the correct "he" (the subject "man" who is hit-

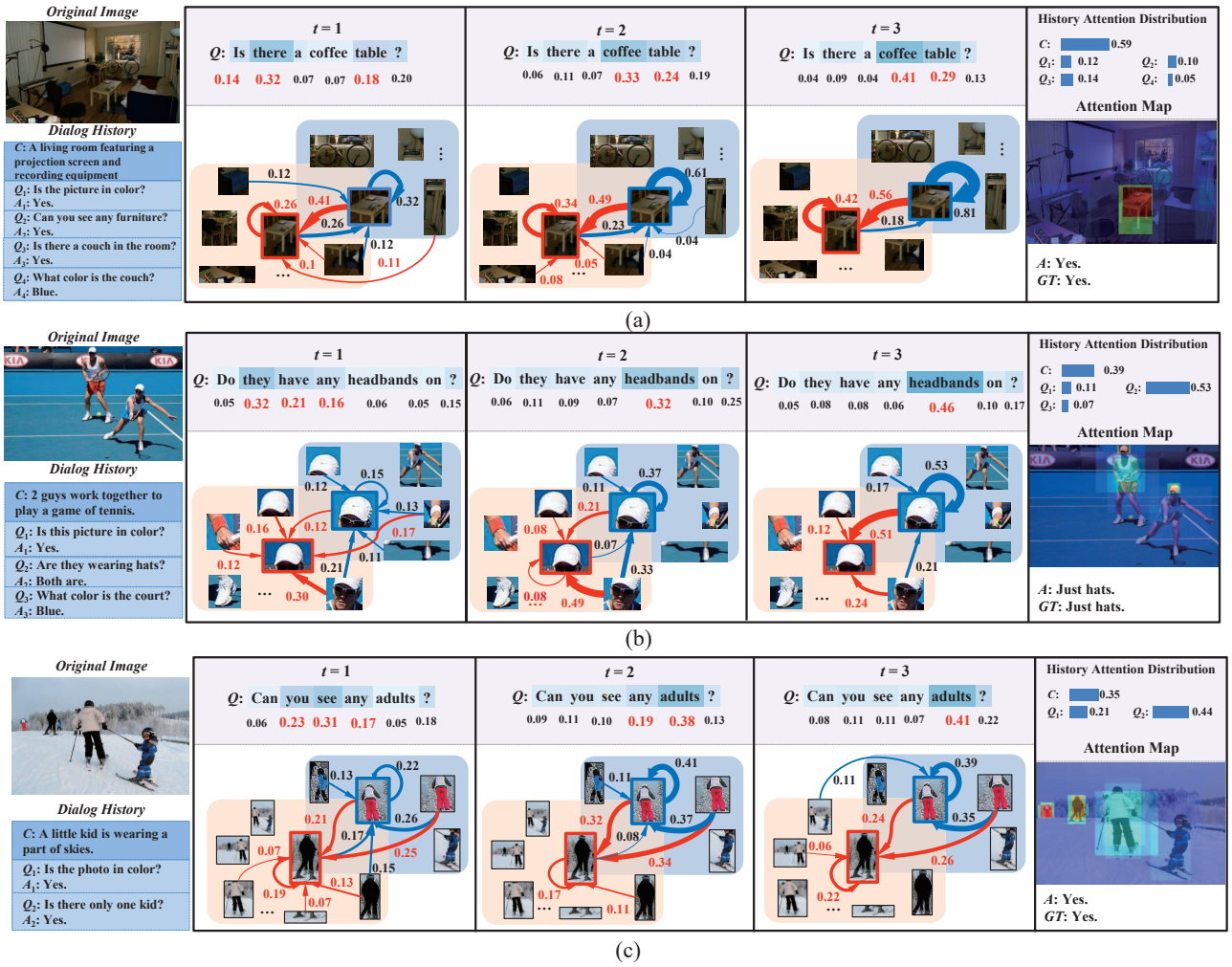

Figure 10. Additional visualization examples on VisDial 0.9. In these attention maps, red and blue bounding boxes correspond to the top-2 attended object nodes in the graph attention learning, respectively.

ting the baseball with the bat in the previous dialogue). Besides, as shown in Fig. 9 (b), although **CAG w/o *u*** infers the correct answer, but we observe that there is much more reasonable inference using **CAG** than **CAG w/o *u***. For the question Q: "Is there other vehicles?", **CAG** does not attend the bus in the center of picture and devote to searching other vehicles, while **CAG w/o *u*** focuses on all the vehicles.

In a nutshell, **CAG w/o *u*** is accustomed to attend all the objects appeared in the question, while **CAG** tries to ground the relevant objects discussed in the entire dialogue. If without the history reference, the dialogue agent can not perform the pronoun explanation (*e.g.*, the visual grounding of "he", "it" and "there", *ect.*) well, and then the subsequent iterative inferences are affected. Therefore, the history-related context u is necessary for the visual-textual co-reference reasoning in our solution.

A.3.2 Additional qualitative results of visual-reference

We provide additional four visualization results in Fig. 10. These qualitative results also demonstrate that CAG has interpretable textual and visual attention distribution, reliable context-aware graph learning, and reasonable inference process.