Supplementary Material for "Making History Matter: History-Advantage Sequence Training for Visual Dialog"

This supplementary document will further detail the following aspects in the main paper:

1) Section 1 details the implementation of HACAN.

2) Section 2 provides more qualitative results.

1. Implementation of HACAN

In this section, we detail the network architecture of our propose HACAN and all modules in the network.

1.1. Gated History-Awareness (GHA).

GHA module first calculates the relationship of the current question and the last round in the history, then provides a history-aware weight to determine whether to reuse the feature from the last round to initialize the first FCA module in HACAN. If the current question refers to the last round, the model may focus on the same as what the last round focuses on, and benefit from the outputs of the last round. The exact architecture of GHA is given in Table 6.

1.2. Feature-Wise Co-Attention (FCA)

FCA module calculates the attention weight with the guidances and aggregates the features in a feature level. The inputs are one input matrix $X \in \mathbb{R}^{N \times d}$ in the input triplets $\{Q^t, H^t, V^t\}$, where N is the number of features and two guidances from GHA or ECA. Three FCA modules are applied in parallel. The exact architecture of FCA is given in Table 7.

1.3. Element-Wise Co-Attention (ECA)

ECA module is designed as a attention module in an element level. It activates the relevant elements and restrains the irrelevant ones with the guidances. The inputs are the outputs of three FCA modules. The exact architecture of ECA is given in Table 8.

1.4. Residual Connection

In HACAN, each FCA and ECA are bundled together as an attention layer. The attention layers are applied in a residual manner after GHA. The exact architecture of HACAN with 2 attention layers is given in Table 9.

Table 6. Architecture of Gated History-Awareness.

Index	Input	Operation	Output Shape
(1)	$\boldsymbol{Q}^t$	Input Feature	$K \times d$
(2)	$\boldsymbol{h}_e^{t-1}$, $\boldsymbol{v}_e^{t-1}$	From last round	d, d
(3)	(1)	Self Attention	d
(4)	(2), (3)	Linear ($d \to d$)	d, d
(5)	(4)	Concatenate	$2d$
(6)	(5)	MLP ($2d \to 1$)	1
(7)	(6)	Sigmoid	1
(8)	(2), (7)	Weighted Output	d, d

Table 7. Architecture of Feature-Wise Co-Attention.

Index	Input	Operation	Output Shape
(1)		Input Feature	$N \times d$
(2)		From GHA/ECA	d, d
(3)	(2)	Expand as (1)	$N \times d$, $N \times d$
(4)	(1), (3)	Linear ($d \to d$), Sum	$N \times d$
(5)	(4)	Tanh	$N \times d$
(6)	(5)	Linear($d \to 1$)	N
(7)	(6)	Softmax	N
(8)	(1), (7)	Weighted Sum	d

Table 8. Architecture of Element-Wise Co-Attention.

Index	Input	Operation	Output Shape
(1)		From FCA	d
(2)		From FCA (guidances)	d, d
(3)	(1)	Outer Product	$m \times d$
(4)	(2)	Expand as (3)	$m \times d$, $m \times d$
(5)	(3), (4)	Linear($d \to d$), Sum	$m \times d$
(6)	(5)	Tanh	$m \times d$
(7)	(6)	Linear ($m \to 1$)	d
(8)	(1), (7)	Hadamard Product	d

2. Additional Qualitative Results

We have conducted ablation study in the main paper to evaluate our proposed method. The provided qualitative re-

Table 9. Architecture of HACAN.

Index	Input	Operation	Output Shape
(1)	Q^t	GHA (Table 6)	d, d
Attention Layer 1 (AL 1)			
(2)	(1), $\boldsymbol{H}^t$	FCA for $\boldsymbol{H}^t$ (Table 7)	d
(3)	(1), $\boldsymbol{Q}^t$	FCA for $\boldsymbol{Q}^t$	d
(4)	(1), $\boldsymbol{V}^t$	FCA for $\boldsymbol{V}^t$	d
(5)	(2), (3), (4)	ECA for (2) (Table 8)	d
(6)	(2), (3), (4)	ECA for (3)	d
(7)	(2), (3), (4)	ECA for (4)	d
Attention Layer 2 (AL 2)			
(8)	(6), (7), $\boldsymbol{H}^t$	FCA for $\boldsymbol{H}^t$	d
(9)	(5), (7), $\boldsymbol{Q}^t$	FCA for $\boldsymbol{Q}^t$	d
(10)	(5), (6), $\boldsymbol{V}^t$	FCA for $\boldsymbol{V}^t$	d
(11)	(8), (9), (10)	Repeat (5)-(7)	d, d, d
(12)	AL 1, AL 2	Residual Connection	d, d, d
Decoder			

sults demonstrate reliable contextual reasoning and history sensitive of our method. In this section, we show more qualitative results in details.

In Figure 5, our model is given an image with persons and a flying kite in the air. The history of the dialog contains several pronouns (*e.g.* "they" in Q3 and Q4, "it" in Q5). We visualize the visual attention of the given image when our model answers Q4 (light orange part) and Q5 (light blue part). Each visualization image in Figure 5 shows the visual attention after one layer. The image of GHA illustrates the weighted visual attention of the last round after GHA. In the history, the ambiguous question Q4 has the same topic with Q3, and two pronouns ("they") both refer to "people" in Q2. The GHA finds the relationship of Q4 and Q3 that they may talk about the same topic and calculates a history-awareness weight (Index (7) in Table 6). As a result, our model determines to reuse the visual feature from Q3 and the history-awareness weight is close to 1. Benefitting from the precise guidance initialization (Index (2) in Table 7), the follow-up attention layers easily focus on the right regions in the image, helping the model generate the response correctly. However, the pronoun "it" doesn't refer to "they", as the attended history in Q4 is mainly about "people" (Index (2) in Table 6), which has nothing to do with "flying" in Q5 (Index (3) in Table 6). Our GHA applies the reasoning process by Index (4)-(7) in Table 6. As a result, our model determines to discard the feature.

In addition, the attention layers in Q5 have to compute the attention without the initialization, and refine the attention feature. As illustrated in the light blue part of Figure 5, the visual attention is more precise when with more attention layers, which demonstrates the advantage of HACAN.

Figure 6 illustrates additional qualitative results when given the "gold" history and the fake history, these results also demonstrate that our method is sensitive to the history and has reliable contextual reasoning.

For Example, as illustrated in Figure 6(a), our model concludes from the "gold" history that "you can only see a small portion of his pants". From this, the model chooses the region at the left lower part in the image. Meanwhile, the round in the history has higher attention score, helping the model rank answer options more correctly. However, the fake history is lacking in effective information and makes the model in loss. As a result, the model chooses wrong region and generates worse response. We find that the response changes a lot when the history round has high attention score.

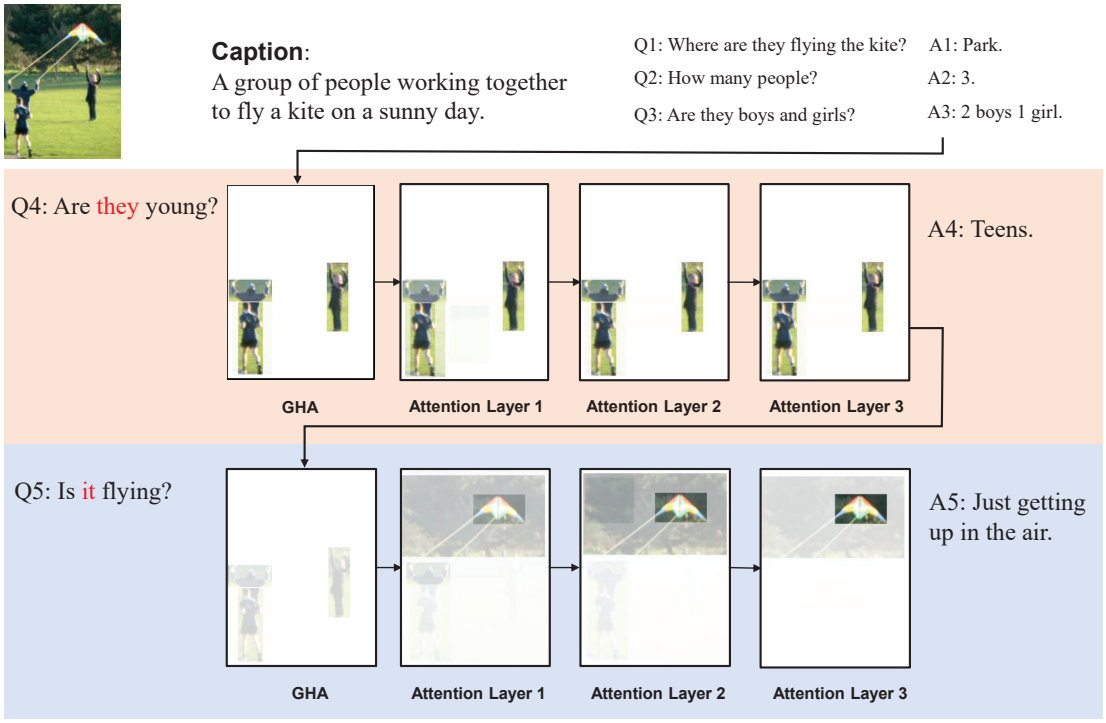

Figure 5. An qualitative result of our method. We visualize the visual attention of the given image when our model answers Q4 (light orange part) and Q5 (light blue part). Each visualization image in the figure shows the visual attention after one layer. The image of GHA illustrates the weighted visual attention of the last round after GHA. As Q4 has the same topic as the topic of Q3, GHA decides to reuse the feature from the 3-rd round as the guidances. On the contrary, GHA in the 5-th discards the feature from the 4-th round. The visual attention is refined by attention layers.

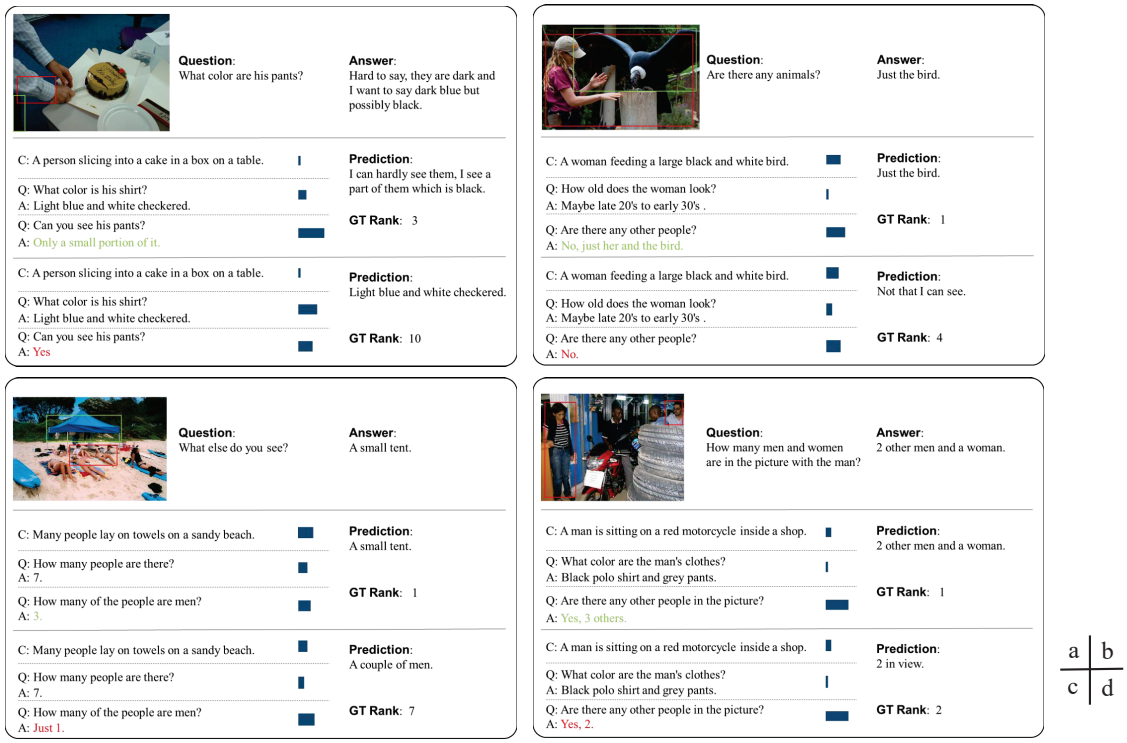

Figure 6. Additional qualitative results of our methods. Incorrect history and the region chosen by model with fake history are marked with red. GT rank denotes the rank of ground-truth answer in the sorted list. The bar graph denotes the history attention. For the sake of aesthetic, we omit the green region in figure (d), as the red region and green region will overlap.