A Reinforcement Learning Algorithm Details

Based on the supervisedly pretrained Q-Bot and A-Bot, RL is used to fine-tune the two models. Following previous works, we use REINFORCE algorithm to update policy parameters.

Q-Bot's policy is formulated as $\pi_Q(q_t|S_t;\theta_Q)$, where S_t is the state, q_t is the generated question and the policy is learned by the deep neural network parameterized by θ_Q. Similarly, A-Bot's policy is formulated as $\pi_A(a_t|S_t;\theta_A)$. For policy π_Q and π_A, the objective function of the policy gradient is given by: $J(\theta_Q,\theta_A) = E_{\pi_Q,\pi_A}\sum_{t=1}^{T} r_t(S_t,A_t)$. Following the REINFORCE, the gradient of $J(\theta_Q)$ can be written as: $\nabla J(\theta_Q) = E_{\pi_Q,\pi_A} r_t(\cdot)\nabla\theta_Q log\pi_Q(q_t|S_t)$. Similarly, the gradient of $J(\theta_A)$ can be written as: $\nabla J(\theta_A) = E_{\pi_Q,\pi_A} r_t(\cdot)\nabla\theta_A log\pi_A(a_t|S_t)$. Accordingly, Q-Bot and A-Bot are optimized with policy gradients while maximizing the reward r_t as claimed in Eq. 8. In Algorithm 1, we exemplify the training procedure that uses RL at the beginning of the self-talking dialogue. To simplify, the target image feature is denoted by tar_feat and the predicted image feature is denoted by $pred\_feat$.

B Training Details

Supervise Learning. We use the same Q-Bot as in previous work [8]. In concrete, the model consists of 4 components: 1) fact encoder, 2) history encoder, 3) question decoder, 4) feature regression network. Following previous settings, all encoders and decoder are 2-layer LSTM with 512-d hidden states, the feature regression network is a 1-layer MLP. Fact encoder and history encoder encodes the textual dialogue history. Based on the encoding of dialogue history, question decoder (a 2-layer LSTM) generates the question and feature regression Network (a single fully connected layer) outputs the predicted image feature. Q-Bot is optimized with cross-entropy loss and L2 loss that minimizes the distance between predicted image feature and target image feature. Following previous works, we pretrain Q-Bot for 20 epochs with a learning rate of 1e-3 that is iteratively decayed to a minimum of 5e-5. Adam is the optimizer and dropout rate is 0.5.

Among many attention-based A-Bot models, we use a classic attention-based A-Bot with History-Conditioned Image Attentive Encoder (HCIAE) [16] to verify the proposed method. According to the visual attention computation method in HCIAE, the current concerned visual contents $\mathbf{i\text{-}vc_t}$ in Eq. 2 can further be described as:

$$\mathbf{att_t} = softmax(\mathbf{w_a}^T tanh(\mathbf{W_H M_t^H} + (\mathbf{W_q m_t^q})\mathbf{1^T}), \tag{9}$$

where $\mathbf{M_t^H}$ is history feature, $\mathbf{m_t^q}$ is question feature, $\mathbf{w_a}$, $\mathbf{W_H}$ and $\mathbf{W_q}$ are learnable parameters. A-Bot is optimized with cross entropy loss. Following their implementation, we pretrain A-Bot for 40 epochs with a learning rate of 4e-4, decayed by 0.75 after 10 epochs. We use Adam as the optimizer. Dropout rate is 0.5.

Reinforcement Learning. The γ in the polarization operation in Eq. 4 is set to be 0.6 empirically. The reward coefficients in Eq. 8 are set as follows: α_O=1, α_D=1e-1, α_I=1e-2. The coefficients vary because different rewards are at different magnitudes. We set the coefficients based on the consideration of balancing task success and dialogue quality. We give further analysis on coefficients setting in Appendix E.

RL smoothly begins according to a curriculum learning setting: SL is used for the first K rounds of dialogue and policy-gradient update works in the left 10 - K rounds. Following [24], we start at K = 9 and gradually anneal it to 4. We train the agents for 30 epochs with a batch size of 32. As learning rate is inconsistent in SL, we also use different settings in

Algorithm 1 Training the Q-Bot and A-bot using REINFORCE with ECS-based rewards.

for Each Update **do**
  # Initialize A-VC:
  $a\text{-}vc_0 \leftarrow 0$
  $i\text{-}vc_0 \sim U(0,1)$
  $\triangle pred\_feat \leftarrow 0$
  for $t = 1$ to T **do**
    $q_t, pred\_feat \leftarrow QBot(H_t)$
    # Obtain I-VC:
    $a_t, att_t \leftarrow ABot(I, H_t, q_t)$
    $i\text{-}vc_t = att_t$
    # Update A-VC:
    $a\text{-}vc_t \leftarrow OR(a\text{-}vc_{t-1}, polarize(i\text{-}vc_t))$
    # Compute rewards:
    $r_t^D \leftarrow D_{KL}(i\text{-}vc_t || i\text{-}vc_{t-1})$
    if $\sum_{j=1}^{N} a\text{-}vc_t^j - a\text{-}vc_{t-1}^j > 0$ **then**
      $r_t^I \leftarrow 1$
    else $r_t^I \leftarrow 0$
    $r_t^O \leftarrow (tar\_feat, pred\_feat)^2 - \triangle pred\_feat$
    $r_t \leftarrow \alpha_O r_t^O + \alpha_D r_t^D + \alpha_I r_t^I$
    $\triangle pred\_feat \leftarrow (tar\_feat, pred\_feat)^2$
    Evaluate $\nabla J(\theta_Q)$ and update QBot
    Evaluate $\nabla J(\theta_A)$ and update ABot
  end for
end for

RL. We optimize the Q-Bot with a learning rate of 1e-3 that is iteratively decayed up to a minimum of 5e-5 while learning rate of A-Bot is initialized with 1e-4 and decayed up to 5e-6. We clamp gradients to [-5, 5] to avoid explosion.

C Additional Qualitative Analysis on A-Bot

We find our A-Bot tends to generate descriptive responses that not only directly answer the question but also describe correlated visual information. We give the selected generated dialogue examples in Fig. 8(a). As can be seen, when is asked questions like *“what else?"*, our A-Bot gives detailed description about visual contents that have not been mentioned yet while the A-Bot optimized by single original reward tediously says *“no"*. As in the top example, our A-Bot constantly adds new information of the undisclosed image by answering *“yes , a glass of water , a coffee cup , a cup of some sort , and another plate"* to *“do you see any other items ?"* and answering *“a cell phone , part of what looks like a UNK , the rest of the paper , and a... "* to *“anything else seen besides the interest ?”*. The descriptive responses help Q-Bot have better knowledge of the image and encourage Q-Bot to generate new questions based on the given detailed answers – as in the bottom example in Fig. 8(a), after knowing that there are tennis court in the background, Q-Bot then asks *“what color is the surface ?"*.

In Fig. 8(b), we give the examples that replace the jointly-trained Q-Bot with the ground-truth questions, to illustrate the adaptability of our A-Bot and make a direct comparison. As can be seen, our A-Bot still performs outstandingly as it tends to give detailed answers that include additional explanations or descriptions moreover the simple yes/no. In the first example, when the question is *"is this in a zoo?"*, our A-Bot answers *"no, it's in the wild"*, which corrects Q-Bot's belief on the undisclosed image, while the basic A-Bot only answers *"i don't think so"*. In the third example, when the question is *"can you see the leaves of the tree?"*, our A-Bot answers *"yes, they are green"* which gives additional information about the related visual contents. Informative and diverse answers like these are more human-like and are helpful to keep users engaged.

D Additional Qualitative Analysis on Comparing Methods

We give additional generated dialogue examples from comparing methods in Fig. 9. As what has been analyzed before, our method helps achieve less repetitive and more visual informative dialogues. As can be seen, dialogues generated by *RL* start to repeat at an early stage of the dialogue, dialogues from *RL-Diverse* are less repetitive but tend to repeat at successive rounds, while dialogues generated by *RL-Ours* can always converse on new visual contents. Besides, our dialogues valuably show consistency sometimes. For example, in the 2nd sample, after the A-Bot answers *"shirt and pants"*, the Q-Bot further asks *"What color is his shirt?"*. Also, in the 3rd sample, as the A-Bot mentions *"there is a candle"*, the Q-Bot then asks *"What color is the candle?"*

E Analysis on Different Settings of Reward-Coefficients

We explore the training effects brought by different settings of reward-coefficients, i.e. the coefficients of DR and IR. We show the results in Tab. 2, which includes indicators that measure the performances in three-aspects: 1) joint performance of Q-Bot and A-Bot by PMR, i.e. the Percentile Mean Rank in the image-guessing task; 2) Q-Bot performance by Unique Question; 3) A-Bot performance by MRR and NDCG.

In setting 1, we intuitively assign the reward coefficients to let weighted rewards be in the same order of magnitude. We also conduct experiments with other reward-coefficients settings, as in 2 and 3. We think different coefficient settings lead to different training results on Q-Bot and A-Bot. Setting 1 achieves the most balanced cooperative Q-Bot and A-Bot while other settings may help improve the performances on separate tasks, i.e., image-guessing and Visual Dialogue answering, which are also meaningful.

settings	α_D	α_I	Unique question^{Q}	MRR^{A}	NDCG^{A}	PMR^{QA}
1	0.1	0.01	**8.11**$_{\pm 0.08}$	49.36$_{\pm 0.00}$	57.81$_{\pm 0.00}$	96.86$_{\pm 0.06}$
2	0.05	0.01	6.79$_{\pm 0.07}$	49.38$_{\pm 0.00}$	**58.00**$_{\pm \mathbf{0.00}}$	96.57$_{\pm 0.08}$
3	0.01	0.001	5.43$_{\pm 0.00}$	**49.42**$_{\pm \mathbf{0.00}}$	57.84$_{\pm 0.00}$	**97.00**$_{\pm \mathbf{0.00}}$

Table 2: Performances on VisDial v1.0 val under different reward-coefficients settings. α_D and α_I are coefficients for DR and IR, respectively. The superscripts represent the indicator belongs to which agent or both.

Image	RL	RL - Ours
	is this in color ? A:yes do you see any people ? A:just a hand is this inside or outside ? A:i ca n't tell it 's a close up what color is the plate ? A:white anything else in the room ? A:no , just the hot dog anything else in the room ? A:no , just the hot dog anything else in the room ? A:no , just the hot dog anything else in the room ? A:no , just the food anything else in the room ? A:no , just the hot dog anything else in the room ? A:no , just the food	is the photo in color ? A:yes is it a professional photo ? A:no is it well lit ? A:yes is it daytime ? A:yes is it sunny ? A:yes do you see any people ? A:yes do you see any other items ? A:yes , a glass of water , a coffee cup , a cup of some sort , and another plate what color is the shirt ? A:i ca n't see the top of the shirt anything else seen besides the interest ? A:a cell phone , part of what looks like a UNK , the rest of the paper , and a what color is the table ? A:i ca n't see the table
	is this in color ? A:yes how many people do you see ? A:2 are they male or female ? A:male how old do they look ? A:30 's anything else in the room ? A:no anything else in the room ? A:no anything else in the room ? A:no anything else in the room ? A:no anything else in the room ? A:no anything else in the room ? A:no	is the photo in color ? A:yes is it a professional photo ? A:no is it well lit ? A:yes is it daytime ? A:yes how old are the men ? A:20 's what race are the men ? A:white what else is in the background ? A:tennis rackets , a tennis court , part of a tennis ball , part of a blue sky what color is the man 's shirt ? A:orange anything else in the background ? A:no , just a baseball glove , part of a tennis ball , part of a tennis player , the what color is the surface ? A:green

(a)

	Question	RL A-Bot	RL-Ours A-Bot
	is this photo in color ?	yes	yes
	is this outdoors ?	yes , it is	yes , it is
	is it sunny ?	***yes , it is***	***no , it 's cloudy***
	can you see any trees ?	no	no
	can you see grass ?	yes , there is a lot of grass	yes , i can see grass
	how many zebras are there ?	there are 2	2
	what color are they ?	black and white	black and white
	can you see any dirt ?	***yes , there is a lot of grass***	***yes , in the background***
	is this in a zoo ?	***i do n't think so***	***no , it 's in the wild***
	is the field grassy ?	yes	yes
	is there any people in the picture ?	no	no
	is there anything in the picture ?	yes	yes
	is it dark ?	no	no
	is it a big room ?	***no***	***no, it 's small***
	what color is the carpet ?	***brown***	***multi colored***
	are the lights big ?	yes	yes
	what color of the towel ?	white	white
	is the vanity large ?	yes	yes
	anything in the sink ?	no	no
	anything else stand out ?	***no***	***yes , there is a cabinet***
	is this photo in color ?	yes	yes
	does the photo look professional ?	no	no
	is it daytime ?	yes	yes
	is it sunny ?	yes	yes
	are there people ?	no	no
	are there animals ?	no	no animals
	can you see any signs ?	***yes***	***yes , but i ca n't read them***
	can you see the leaves of the tree ?	***yes***	***yes , they are green***
	are there any cars ?	no	no cars
	are there any buildings ?	no	no

(b)

Figure 8: Illustration of A-Bot's ability to generate descriptive responses when (a) answering jointly optimized Q-Bot, (b) answering ground truth questions. Comparing models are: 1)'RL': basic agents fine-tuned by RL using the original rewards; 2) RL-Ours: basic agents fine-tuned by RL with two ECS-based rewards in addition.

Image + Caption	RL	RL - Diverse	RL - Ours
 Caption: 2 dogs with a hat standing beside each other	what color are dogs ?1 is white and other is brown are they big dogs ? medium sized what color are they ?1 is white and other is black anything else in room ?there is fence in background anything else in room ?nothing else anything else in room ?nothing else anything else in room ? nothing else anything else in room ?nothing else anything else in room ?nothing else anything else in room ?nothing else	what kind of dogs are they ?they are labs what color are dogs ?1 is white and other is brown what kind of dogs ?they look like labs what color are they ?they are white are they inside or outside ? outside are there any trees around ?yes are there any people in picture ?no are there any trees ?yes are there any buildings ?no are there trees around ?there are trees in background	is the photo in color ?yes is it a professional photo ?no is it well lit ?yes is it daytime ?yes what color are the dogs ?1 is white and 1 is brown are the dogs big or small ?big do you see anything else besides the dogs ?no do you see anything else besides interest ?no what color is the floor ?brown anything else in the background ?no
 Caption: a man stands at a dining room table holding a knife to cut a cake	what color is his hair ?dark brown is he wearing glasses ?no what color is his hair ?dark brown is he wearing glasses ?no anything else in the room ?yes , a lot of glasses , coffee table anything else in the room ?there is a small table in the background , and part of the table , part of the table , anything else in the room ?there is a small table in the corner of the table , part of the table , part of the anything else in the room ?there is a small table behind the man , and the man 's hand is on the wall behind the anything else in the room ?there is a small table in the background , and the corner of the man , the man 's hand anything else in the room ?there is a glass , a table , a table	how old is the man ?late 20 's early 30 's what is he wearing ?a button up shirt and a tie what color is his hair ?brown is he wearing glasses ?no what color is his hair ?brown what color is his shirt ?white what color is the table ?brown can you see furniture around ?yes , a table , a table can you see furniture around ?yes , there is a table , a table can you see furniture around ?yes , there is a table , a table	is the photo in color ? yes is the photo close up ? no how old is the man ?20 's what race is the man ? white how old does he look ? 30 what is he wearing ?shirt and pants is there anything else other than him ?yes , a bookshelf , shelves what color is his shirt ? white anything else in the background ?tables , tables , tables what color is the floor ?i ca n't see the floor
 Caption: an elderly woman posing with 2 girls and a boy in front of a birthday cake in the kitchen	how old does the woman look ?in her 30s what is she wearing ?a white shirt what kind of cake is it ?white cake how many candles are on the cake ?0 anything else in the room ?yes , a table , cups what color are the cups ?clear glass what color are the cups ?clear glass anything else in the room ?there are a few decorations in the background , but i ca n't tell what they are anything else in the room ?there are a few decorations in the background , but i ca n't tell what they are anything else in the room ?there is a table behind the cake , and the table is in the background , but i can	is the photo in color ?yes how old is the girl ?maybe 5 what color is her hair ?brown what is she wearing ?a white shirt what color is her hair ?brown what is she wearing ?a white shirt what color is her hair ?brown what color are the walls ?white can you see furniture around ?yes , part of a table , a table can you see furniture around ?yes , the kitchen table and the kitchen is in the kitchen , the kitchen is in the kitchen	how old is the woman ?i would say late 20 's early 30's how old is the woman ? she looks like she 's in her 20 's what color is her hair ? her hair is brown what color is the cake ? the cake is white what color is the girls hair ?her hair is brown is there anything else on the table ?yes , there is a candle and a candle what color is the candle ?the candles is yellow is there anything else in the background ?yes , there are a lot of stuff in the background ,cabinets what color is the floor ? the floor is not visible anything else in the room ?there are a lot of things on the table , and the cabinets are the kitchen cabinets , the other
 Caption: a large america airlines plane flying through the air	what color is the plane ?red , white , and blue is there any writing on it ?yes what does it say ?UNK what color is the plane ?red , white , and blue is there any writing on the plane ?yes what does it say ?UNK anything else in the background ?no anything else in the room ?no anything else in the room ?no anything else in the room ?no	is the photo in color ?yes is there any people ?no is this indoors or outdoors ?outdoors can you see the sky ?yes can you see the clouds ?no can you see trees ?no can you see buildings ?no can you see trees around ?no can you see buildings around ?no can you see trees around ?no	what color is the plane ? red and white is it a passenger plane ? yes can you see the windows on the plane ? yes can you see the windows on the plane ? yes can you see any people ?no can you see the sky ?yes can you see anything else besides interest ?no what color is the plane ? white and red can you see anything else besides the plane ? no can you see the ground ?no

Figure 9: Additional generated dialogue Examples from comparing methods, i.e., 1)‘RL’: basic agents fine-tuned by RL using the original rewards; 2)‘RL-Diverse’: diverse agents fine-tuned by RL using the original rewards; 3) RL-Ours: basic agents fine-tuned by RL with ECS-based rewards in addition.