The Dialog Must Go On:
Improving Visual Dialog via Generative Self-Training
(Supplementary Materials)

Gi-Cheon Kang[1,2] Sungdong Kim[3*] Jin-Hwa Kim[3,2*] Donghyun Kwak[4*] Byoung-Tak Zhang[1,2]
[1]IPAI, Seoul National University [2]AIIS [3]NAVER AI Lab [4]NAVER Cloud CLOVA
{chonkang, btzhang}@snu.ac.kr {sungdong.kim, j1nhwa.kim, donghyun.kwak}@navercorp.com

Overview. The supplementary materials are organized as:

- Section A shows a detailed model architecture.
- Section B presents further quantitative analysis.
- Section C presents further qualitative analysis.
- Section D presents implementation details.
- Section E shows discussion, including limitations, future work, and ethical considerations.

A. Details of model architecture

A detailed architecture of our proposed model is presented in Figure 1. We use the encoder-decoder model, where the encoder aggregates the multimodal context, and the decoder generates the target sentence using the hidden states of the encoder. The answerer models (*i.e.,* the student and the teacher) utilize the given image, the dialog history, and the question as the context. On the other hand, the questioner uses the given image and the dialog history as context to generate the question.

We employ the ViLBERT model [23] as our encoder. We employ the $\text{BERT}_{\text{BASE}}$ model [7] for sequence generation [31] as our autoregressive decoder. The decoder has 12 layers of transformer blocks, with each block having 12 attention heads and a hidden size of 768. We present a detailed view of the encoder in (b) for Figure 1. The encoder consists of the vision stream and the language stream. The language stream is the same model as the decoder (*i.e.,* $\text{BERT}_{\text{BASE}}$), which has 12 layers of transformer blocks. The vision stream has 6 layers of transformer blocks, with each block having 8 attention heads with a hidden size of 1024. The co-attention layers connect the 6 transformer layers in the vision stream to the last 6 transformer layers in the language stream. The encoder concatenates the hidden states of each stream and passes them to the decoder. The decoder generates the target sentence by using them.

*Equal contribution

B. Further quantitative analysis

B.1. Experiments on the discriminative models

In this subsection, we discuss the details regarding GST for the discriminative visual dialog. We first describe how we can adapt GST to the discriminative models and then show the results on VisDial v1.0 test-standard split.

Model architecture. Although our main focus is the generative model, we conduct additional experiments to identify the effect of GST in the discriminative VisDial model. Our proposed models (*i.e.,* the student, the teacher, and the questioner) are based on encoder-decoder architecture where the encoder is based on the vision-and-language encoder model [23], and the decoder is the transformer decoder [31]. In this experiment, we remove the decoder model, so the student is based on the encoder-only architecture, the same model architecture as the ViLBERT model [23]. We describe more details in the following subsection.

Tricks for adapting to a discriminative task. The goal of the discriminative task is to retrieve the ground-truth answer from a list of answer candidates. Therefore, it implies that the gold VisDial dataset [6] contains the pre-defined answer candidates for each question to train and evaluate the discriminative models. However, the silver VisDial dataset generated by our proposed models does not include the answer candidates since the dataset is generated to train the generative models that do not need the answer candidates. To circumvent this issue, GST first trains the student model for the generative task, *i.e.,* the encoder-decoder model, on the silver VisDial data. Then, we extract the trained weights of the encoder in the student and initialize the encoder-only model with the weights. Finally, the encoder-only model is trained to retrieve the ground-truth answer from the list of answer candidates using the gold VisDial dataset. This trick circumvents the need for the answer candidates when training the silver VisDial data.

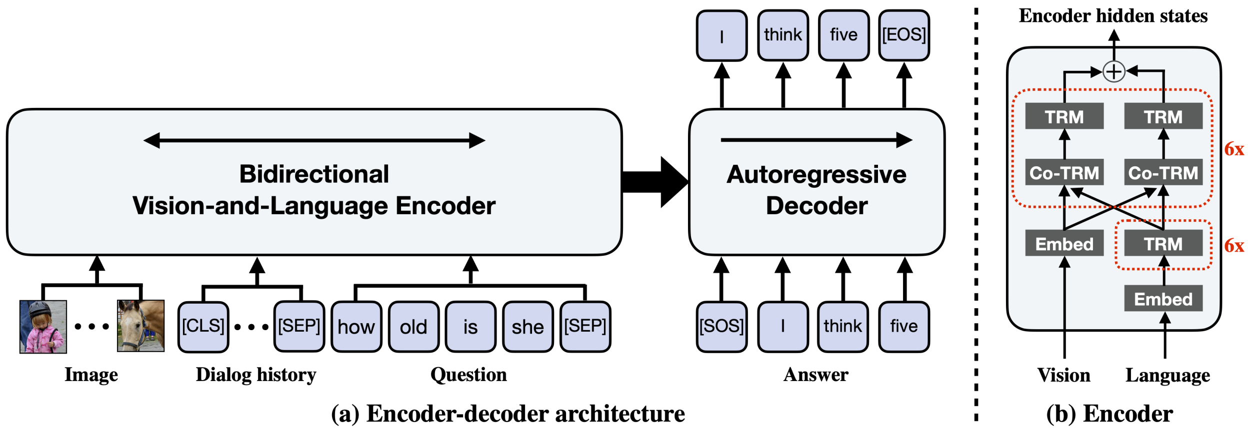

Figure 1. A detailed architecture of our proposed model. We propose the encoder-decoder model where the encoder aggregates the given multimodal context, and the decoder generates the target sentence. (b): a more detailed view of the encoder. TRM and Co-TRM denote the transformer module and the co-attentional transformer module, respectively. ⊕ denotes the concatenation operation.

	VisDial v1.0 (test-std)					
Model	NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
CorefNMN [17]	54.70	61.50	47.55	78.10	88.80	4.40
RvA [25]	55.59	63.03	49.03	80.40	89.83	4.18
Synergistic [11]	57.32	62.20	47.90	80.43	89.95	4.17
ReDAN [9]	61.86	53.13	41.38	66.07	74.50	8.91
DAN [13]	57.59	63.20	49.63	79.75	89.35	4.30
FGA [34]	52.10	63.70	49.58	80.97	88.55	4.51
VD-BERT [39]	59.96	65.44	51.63	82.23	90.68	3.90
VisDial-BERT [24]	63.87	67.50	53.85	84.68	93.25	3.32
Student (ours)	**64.91**	**68.44**	**55.05**	**85.18**	**93.35**	**3.23**
P1+P2† [28]	71.60	48.58	35.98	62.08	77.23	7.48
MCA† [1]	72.47	37.68	20.67	56.67	72.12	8.89
SGL+KT† [14]	72.60	58.01	46.20	71.01	83.20	5.85
VD-BERT† [39]	**74.54**	46.72	33.15	61.58	77.15	7.18
UTC† [5]	74.32	50.24	37.12	63.98	79.88	6.48
VisDial-BERT† [24]	74.47	50.74	37.95	64.13	80.00	6.28
Student† (ours)	71.76	**68.09**	**55.18**	**83.68**	**91.93**	**3.57**

Table 1. Test-std performance of the discriminative model on the VisDial v1.0 dataset. ↑ indicates higher is better. ↓ indicates lower is better. † denotes the use of dense labels.

Results on VisDial v1.0 test split. We compare the student model with the state-of-the-art approaches in the discriminative task, consisting of VisDial-BERT [24], UTC [5], VD-BERT [39], SGL+KT [14], P1+P2 [28], MCA [1], FGA [34], ReDAN [9], DAN [13], Synergistic [11], RvA [25], and CorefNMN [17]. As shown in the upper part of Table 1, GST outperforms the state-of-the-art approaches on all evaluation metrics in the VisDial v1.0 test-standard split. It is worth noticing that GST boosts NDCG 1.04% (63.87 → 64.91) compared with the VisDial-BERT model, whose configuration is almost the same as the student except for the use of the silver VisDial data. Furthermore, recent studies finetune the discriminative VisDial models on the densely annotated labels[1] in the validation dataset and evaluate the models on the test set to boost NDCG. The dense annotation finetuning yields considerable improvements on NDCG and counter-effect on other metrics (*i.e.,* MRR, R@k, and Mean) due to the trade-off relationship [24] between NDCG and the others. To mitigate such performance polarization, we follow the knowledge transfer technique in SGL+KT [14] when using the dense labels. In the below part of Table 1, the student model still shows competitive performance on NDCG, maintaining powerful performance on other metrics.

[1]https://visualdialog.org/challenge/2019#evaluation

Model	PPL	MCR	IIR	Iteration	VisDial v1.0 (val)					
					NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
Teacher				0	64.50	52.06	42.04	62.92	71.06	14.54
Teacher (w/ CPT)			✓	0	63.59	51.70	41.99	61.88	68.62	16.21
Student (iter1, w/o PPL)		✓	✓	1	63.96	52.33	42.68	62.52	69.47	15.56
Student (iter1, w/o MCR)	✓		✓	1	63.71	52.49	42.56	62.87	70.00	15.21
Student (iter1, w/o IIR)	✓	✓		1	64.57	52.33	42.10	63.46	**71.54**	**14.31**
Student (iter1)	✓	✓	✓	1	65.06	52.84	42.74	<u>63.66</u>	71.30	14.60
Student (iter2)	✓	✓	✓	2	<u>65.46</u>	<u>53.04</u>	**43.15**	63.63	71.00	14.73
Student (iter3)	✓	✓	✓	3	**65.47**	**53.19**	<u>43.08</u>	**64.09**	<u>71.51</u>	<u>14.34</u>

Table 2. Ablation study on the VisDial v1.0 validation split. CPT denotes continued pre-training.

Model	Pre-train # Images	VisDial v1.0 (val)	
		NDCG↑	MRR↑
BLIP [19]	129M	-	69.41
Student (ours)	6.7M	65.92	**69.51**

Table 3. Comparison with BLIP [19] on the VisDial v1.0 validation split. The Pre-train # Images denotes the number of utilized images before finetuning on the VisDial v1.0 data.

Results on VisDial v1.0 validation split. We also compare GST with the state-of-the-art vision-and-language pre-training model, BLIP [19]. The BLIP model is trained on the large-scale image-text datasets, such as Laion-400M [33], CC12M [4], CC3M [37], COCO [22], Visual Genome [18], and SBU captions [26]. Then, the model is finally fine-tuned on VisDial data. GST trains the student model on nearly 6.7M images, including 3.1M images (CC3M [37] and VQA [3]) to pretrain ViLBERT [23] and 3.6M images filtered from CC12M [4] to generate and train synthetic dialog data. As shown in Table 3, GST shows competitive performance on the VisDial v1.0 validation split, outperforming BLIP on MRR. It is noticeable that the BLIP model utilizes nearly twenty times more images than GST. It indicates that GST is effective and sample-efficient.

B.2. Ablation study

We perform an ablation study to illustrate the effect of each component in GST. We report the performance of four ablative models: student w/o PPL, student w/o MCR, student w/o IIR, and teacher w/ CPT. Student w/o PPL denotes the model that utilizes all generated QA pairs without applying the perplexity-based data selection. Student w/o MCR does not inject noises into the inputs of the student model. Student w/o IIR utilizes the entire CC12M [4] images to generate the silver VisDial data without applying in-domain image retrieval. It is the same model as the student-iter1-full in Section 4.3. Lastly, the teacher with continued pre-training (CPT) continues to perform pre-training with image-caption pairs in the silver VisDial data. CPT is proposed to identify the effect of utilizing additional vision-and-language data. Specifically, masked language modeling loss and masked image region loss are optimized by following ViLBERT [23].

In Table 2, we observe all components (*i.e.,* PPL, MCR, and IIR) play a significant role in boosting the performance. Notably, by comparing the student model with the student w/o IIR, we find that utilizing the entire Web images does not contribute to an accurate answer prediction. Moreover, we observe that CPT results in a considerable drop in performance. We conjecture that it is due to low-precision image captions in the CC12M dataset, as mentioned in the paper [4]. But the student still shows competitive performance even if it also utilizes the captions in the dialog history. Finally, the iterative training monotonically improves the performance, similar to the robustness results in Section 4.3.

B.3. Do performance improvements come from a larger computational cost?

It takes more computational costs to train the student model than to train the teacher model due to the silver VisDial data. Accordingly, we perform an analysis to prove that the performance improvements do not merely come from larger computational costs. The training time of the teacher model is about 1 day with one NVIDIA A100 GPU. It takes 5 days to train the student model with three iterations (*i.e.,* iter3). Accordingly, we compare the ensemble of 5 teacher models with the student model with the iter3. We ensemble 5 teacher models with different weight initialization and average logits for 5 teacher models to predict the answer. The results are shown in Table 4. The student model outperforms the ensembles of 5 teacher models on both metrics. It indicates that the improvements from GST do not merely come from increased computational costs.

B.4. The QA utilization across different iterations

We identify how many QA pairs in the silver VisDial data are actually utilized after applying perplexity-based data selection (*i.e.,* PPL). Accordingly, we define QA utilization as the proportion of utilized QA pairs in the silver VisDial data. The QA utilization across different iterations is shown

Model	VisDial v1.0 (val)	
	NDCG↑	MRR↑
Teacher (single model)	64.50	52.06
Teacher (5 ensembles)	64.82	52.51
Student (single model)	**65.47**	**53.19**

Table 4. Comparison between the student model with the ensemble of the five teacher models on balanced computational costs.

Model	QA Utilization
Student (iter1)	32.52%
Student (iter2)	39.06%
Student (iter3)	**46.40%**

Table 5. We define QA utilization as the proportion of utilized QA pairs in the silver VisDial data after applying perplexity-based data selection (*i.e.,* PPL). The selection threshold τ is fixed at 50.

in Table 5. We observe that the QA utilization increases as the iteration proceeds. It implies that the student model leverages more data as the iteration proceeds, and more importantly, the average perplexity of the generated answers gradually decreases. We argue that the drop of the answer perplexity is closely related to the student model being more confident and remaining low-entropy [10, 38].

C. Further qualitative analysis

C.1. More visualization of silver data

We visualize more silver data based on the image-caption pairs in the Conceptual Captions (CC12M) [4] dataset. As shown in Figure 3, the questioner and the student models generate diverse and correct visual dialog data, although the image caption data is noisy. For instance, the image caption in the fourth example (*i.e., Luckily the woman s daughter adopted a puppy from litter so that poppy can keep in touch with it*) is not well grounded with the given image. Still, our proposed models produce the *visually-grounded* QA samples. Finally, the student sometimes fails to generate correct answers (the red-colored text), similar to Figure 2.

C.2. Analysis of silver and gold answers.

We visualize the ground-truth answer (*i.e.,* the gold answer) and the answer predictions from the student and the teacher models given the same context. As shown in Figure 2, the student model indeed produces correct answers compared with the teacher model. Moreover, both models produce many correct or plausible answers, although the predicted answers differ from the gold answers (see the blue-colored text). For instance, for the last question in the third example (*i.e., Is she wearing a bathing suit?*), the student answers "wetsuit" to the question, although the ground-truth answer is "no". We conjecture that the ability to generate such different yet correct answers is evaluated as a high NDCG performance; NDCG considers all relevant responses in the answer candidates.

D. Implementation details

We integrate the vision-and-language encoder [23] with the transformer decoder for sequence generation (*i.e.,* $\text{BERT}_{\text{BASE}}$ [31]) to train the teacher, the questioner, and the student. The decoder has 12 layers of transformer blocks, with each block having 12 attention heads and a hidden size of 768. The maximum sequence length of the encoder and the decoder is 256 and 25, respectively. We extract the feature vectors of the input images by using the Faster R-CNN [2, 30] pre-trained on Visual Genome [18]. The number of bounding boxes for each image is fixed to 36. We set the threshold for PPL τ to 50. We train on one A100 GPU with a batch size of 72 for 70 epochs. Training time takes about 3 days. We use the Adam optimizer [15] with an initial learning rate 1e-5. The learning rate is warmed up to 2e-5 until 10k iterations and linearly decays to 1e-5. In visually-grounded dialog generation, the questioner and the teacher decode the sequences using the top-k sampling [8, 12, 29] with $k = 7$ and the temperature of 0.7. We use the top-k sampling since its computation is cheap yielding accurate and diverse sequences. Furthermore, we apply the 4-gram penalty [16, 27] when generating visual questions to ensure that no 4-gram appears twice in the questions for each dialog.

E. Discussion

E.1. Relationship between self-supervised pre-training and generative self-training.

We develop the teacher, the questioner, and the student models on top of ViLBERT [23] which leverages vision-and-language pre-training. Thus, the teacher can be understood as a typical model that follows the pretrain-then-transfer learning strategy mentioned in the introduction, whereas the student leverages both pre-training and generative self-training. By comparing the student with the teacher, we identify that self-supervised pre-training and GST are complementary modeling capabilities.

E.2. Limitations and future work.

One of the major limitations of our approach is the learning efficiency of the student model. We demonstrate the effectiveness of our proposed method, but there can be more efficient ways to improve the visual dialog model. For example, our method generates the dialog data without considering the difficulty of the question. We believe that the competency-aware or curriculum-based visual dialog generation can make our proposed self-training algorithm more efficient and powerful. We will leave it as a future work.

Image & Caption	Question	GT Answer	Student Model	Teacher Model
	Q1 : How many children are there?	3	3	4
	Q2 : What is the time of the day?	I can't tell that	Daytime	Daytime
	Q3 : Can you see a soccer ball?	Yes	Yes	Yes
	Q4 : What is the children wearing?	Soccer shirts and pants	Uniforms	Uniforms
	Q5 : What is the color of the ball?	White and red	White and red	White with red
	Q6 : Can you see a spectator?	No	No	No
	Q7 : Are they happy enjoying?	Yes	Yes	Yes
The 2 children's teams are playing soccer against each other	Q8 : What is the color of the hair or they are wearing a helmet?	No	Brown	They all have brown hair
	Q9 : Can you see the sky?	No	No	No
	Q10 : Is there any female?	No	No	No
	Q1 : Is there any snow on the ground?	Yes	Yes there is	Yes tons
	Q2 : Are they dressed for skiing?	Yes	Yes	Yes
	Q3 : Color shirt is the woman wearing?	Blue and white	Blue and white jacket	Black
	Q4 : Is there any trees around?	There are	Yes	Yes
	Q5 : What kind of day is it?	A snowy day	Sunny	Sunny
	Q6 : Are the man taller than the woman?	Yes, they are	Yes	Yes
	Q7 : How many hats are in the picture?	3	3 hats	2
2 men and a woman are on slopes wearing their skies	Q8 : Are they all wearing gloves?	Yes, they are all	Yes	Only two of them are
	Q9 : Do they seem to be enjoy the day?	Yes	Yes, they do	Yes, they do
	Q10 : What time of day is it?	Morning	Daytime	Noon
	Q1 : Is the picture in color?	Yes	Yes	Yes
	Q2 : Are they the only ones in the water?	No	No	No
	Q3 : Can you see the beach?	Yes	Yes	Yes
	Q4 : Are there people on the beach?	Yes	Yes	Yes
	Q5 : How old is the man?	20's	20's	20's
	Q6 : What color is his hair?	Brown	Brown	Black
	Q7 : How many hats are in the picture?	Early 20's	20s	Same
A man is teaching a woman how to ride a surfboard	Q8 : What color is her hair?	Brown	Dark brown	Brown
	Q9 : What color is the surfboard?	White	White	White
	Q10 : Is she wearing a bathing suit?	No	Wetsuit	Yes
	Q1 : Is the man wearing a helmet?	No	No	No
	Q2 : Is this outdoors?	Yes	Yes	Yes
	Q3 : Is it a sunny day?	Yes	Yes	Yes
	Q4 : Is the man wearing a hat?	No	No	No
	Q5 : Can you see graffiti?	No	No	No
	Q6 : Is he wearing elbow pads?	No	No	No
	Q7 : Is he wearing shorts?	No	No	No, jeans
	Q8 : Can you see anyone else?	Yes	Yes	Yes, but they are blurry
A young man performs a trick on a skateboard	Q9 : What color is his board?	Black	Black	Black
	Q10 : What color is his shirt?	White with brown spots	White	White
	Q1 : Is it daytime?	Yes	Yes	Yes
	Q2 : Is the sun out?	Yes	Yes	Yes
	Q3 : How old is the woman?	24	20s	Late twenties
	Q4 : Is she alone?	Yes	Yes	Yes
	Q5 : Is she wearing sunglasses?	No	No	No
	Q6 : Is she carrying anything else with her?	No	A bag	A purse
	Q7 : Is she wearing sneakers?	Yes	Yes	No
A woman rides her bike on a sidewalk next to a fence	Q8 : What color is her shirt?	Wearing a black jacket	Black	Black
	Q9 : Is she smiling?	No	Yes	No
	Q10 : Where is she going?	West on the road	Not sure	She's at a park
	Q1 : Is the man young?	I'd say early 20s	Yes	Yes
	Q2 : Is he dressed casual?	Yes, he is	Yes, t-shirt and jeans	Yes, jeans, a polo shirts
	Q3 : Is his hair short or longish?	It is shoulder length	It is medium length	It's pretty short
	Q4 : What color is his shirt?	It is medium brown	Navy blue	Blue with white writing
	Q5 : Is he wearing jeans?	Yes, he is	Yes, he is	Yes, he is
	Q6 : Is this in the city?	Looks like a suburb area	Looks like it	Yes, it looks like it
	Q7 : Are there any trees around?	Yes, 1 tree	A few in the background	Yes, I can see one
	Q8 : How about cars?	No cars	I don't see any cars	No cars are in the photo
A man walks down the street, pass a yellow fire hydrant	Q9 : Are there any other people?	No, he is alone	No other people	No
	Q10 : Is it sunny?	Yes, it is	I can't see the sky, but it is daytime	I can't really tell

Figure 2. A visualization of answer predictions from the student and the teacher model. The red-colored text denotes an incorrect answer. The blue-colored text indicates the prediction different from the ground-truth answer, but it seems correct or plausible.

Image	Machine Conversation (Silver VisDial)
	C : One of the bedrooms at basement level D1 : Can you see the bed? Yes D2 : Is it made? Yes D3 : What color is the bedspread? White and blue D4 : Are there any pillows? Can't see any D5 : Is there a window? Yes D6 : Are there curtains on the window? No they are open D7 : Are there curtains? There are no curtains D8 : Can you see out of the window? Yes I can see some trees D9 : Can you see any people? No people no D10 : Can you tell what kind of trees? No they are very small and green
	C : Cars at the intersection of FM 1960 and Jones Road in northwest Harris County D1 : Are there any buses? No D2 : Are there any trees? No D3 : How many cars are there? More than 10 D4 : Are there any people? No D5 : Are there any animals? No D6 : Can you see the sky? Yes D7 : What color is the sky? Gray D8 : Is it cloudy? Yes it's cloudy D9 : Can you see any buildings? Yes about 4 buildings D10 : Is it night or day? Daytime
	C : Coffee Bay: The green Wild Coast-line D1 : Is there a boat in the water? No D2 : Is the water calm? No, it looks choppy D3 : Are there any animals? I can't see any D4 : Are there any people in the image? No people D5 : Are there any buildings? No, it's a hill and there are no buildings D6 : Are there any vehicles? No vehicles, it's a mountain D7 : Is the mountain covered in snow? No, it's a very green mountain D8 : Are there any trees? No trees, just grass D9 : Is it sunny? **It's very sunny** D10 : Are there any rocks in the grass? Yes, there are rocks
	C : Luckily the woman s daughter adopted a puppy from litter so that poppy \ can keep in touch with it D1 : Is this in color? Yes D2 : Is the photo well lit? Yes D3 : Is this a close up? Yes D4 : What color is the puppy? Brown D5 : What color is her leash? She doesn't has a leash D6 : What color is woman's hair? I can only see her hand D7 : Is this a professional photo? No it's not D8 : Is this a picture showing the baby's parents? **No it's an adult mother** D9 : Are they outside? No they are indoors D10 : Is this a living room? I can't tell
	C : Smiled beautiful mulatto woman is preparing to make a meal from fresh \ vegetables on the table on the modern kitchen D1 : How old does the woman appear to be? In her twenties D2 : Is she the only person visible in the photo? Yes D3 : What color is the kitchen counter? Tan D4 : Can you tell what type of food she is making? No D5 : Is the kitchen clean? Yes D6 : What color is her hair? Black D7 : Is she using a cutting board? Yes D8 : What color is it? Wood D9 : Does she appear to be making any type of vegetables? Yes she is cooking D10 : What kind of vegetables are in the kitchen? Tomatoes

Figure 3. A visualization of the silver VisDial data based on the image-caption pairs in the Conceptual Captions 12M (CC12M) [4] dataset. The red-colored text denotes an incorrect answer.

E.3. Ethical considerations.

Since GST generates the visually-grounded dialogs, our proposed models have the potential to produce biased and offensive language, although arguably to a lesser extent than the open-domain dialog [20,21,32,35,36,40]. We attempt to mitigate ethical concerns such as biases against people of a certain gender, race, age, and ethnicity or the use of offensive content. Our proposed method utilizes the images and the captions in the Conceptual 12M dataset [4], where several data cleansing processes (*e.g.,* the offensive content filtering or replacing each person name with the special <PERSON> token) have been conducted. At least, we could not find any conversation violating the ethical considerations in a manual inspection by visualizing ∼100 synthetic dialogs.

References

[1] Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. History for visual dialog: Do we really need it? In *ACL*, 2020. 2

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018. 4

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 3

[4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 3, 4, 6

[5] Cheng Chen, Yudong Zhu, Zhenshan Tan, Qingrong Cheng, Xin Jiang, Qun Liu, and Xiaodong Gu. Utc: A unified transformer with inter-task contrastive learning for visual dialog. In *CVPR*, 2022. 2

[6] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *CVPR*, 2017. 1

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 1

[8] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *ACL*, 2018. 4

[9] Zhe Gan, Yu Cheng, Ahmed EI Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. Multi-step reasoning via recurrent dual attention for visual dialog. In *ACL*, 2019. 2

[10] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *NIPS*, 2005. 4

[11] Dalu Guo, Chang Xu, and Dacheng Tao. Image-question-answer synergistic network for visual dialog. In *CVPR*, 2019. 2

[12] Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. Learning to write with cooperative discriminators. In *ACL*, 2018. 4

[13] Gi-Cheon Kang, Jaeseo Lim, and Byoung-Tak Zhang. Dual attention networks for visual reference resolution in visual dialog. In *EMNLP*, 2019. 2

[14] Gi-Cheon Kang, Junseok Park, Hwaran Lee, Byoung-Tak Zhang, and Jin-Hwa Kim. Reasoning visual dialog with sparse graph learning and knowledge transfer. In *EMNLP*, 2021. 2

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 4

[16] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. Opennmt: Open-source toolkit for neural machine translation. In *ACL*, 2017. 4

[17] Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Visual coreference resolution in visual dialog using neural module networks. In *ECCV*, 2018. 2

[18] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *ICCV*, 2017. 3, 4

[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 3

[20] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. In *EMNLP*, 2016. 6

[21] Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. Adversarial learning for neural dialogue generation. In *EMNLP*, 2017. 6

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 3

[23] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 1, 3, 4

[24] Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In *ECCV*, 2020. 2

[25] Yulei Niu, Hanwang Zhang, Manli Zhang, Jianhong Zhang, Zhiwu Lu, and Ji-Rong Wen. Recursive visual attention in visual dialog. In *CVPR*, 2019. 2

[26] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011. 3

[27] Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. In *ICLR*, 2018. 4

[28] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. Two causal principles for improving visual dialog. In *CVPR*, 2020. 2

[29] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 4

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 4

[31] Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. Leveraging pre-trained checkpoints for sequence generation tasks. In *Transactions of the Association for Computational Linguistics*, 2020. 1, 4

[32] Abdelrhman Saleh, Natasha Jaques, Asma Ghandeharioun, Judy Shen, and Rosalind Picard. Hierarchical reinforcement learning for open-domain dialog. In *AAAI*, 2020. 6

[33] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 3

[34] Idan Schwartz, Seunghak Yu, Tamir Hazan, and Alexander G Schwing. Factor graph attention. In *CVPR*, 2019. 2

[35] Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, 2017. 6

[36] Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. In *ACL*, 2015. 6

[37] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 3

[38] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020. 4

[39] Yue Wang, Shafiq Joty, Michael R Lyu, Irwin King, Caiming Xiong, and Steven CH Hoi. Vd-bert: A unified vision and dialog transformer with bert. In *EMNLP*, 2020. 2

[40] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. In *ACL*, 2020. 6