The Dialog Must Go On: Improving Visual Dialog via Generative Self-Training

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

Abstract

Visual dialog (VisDial) is a task of answering a sequence of questions grounded in an image, using the dialog history as context. Prior work has trained the dialog agents solely on VisDial data via supervised learning or leveraged pre-training on related vision-and-language datasets. This paper presents a semi-supervised learning approach for visually-grounded dialog, called Generative Self-Training (GST), to leverage unlabeled images on the Web. Specifically, GST first retrieves in-domain images through out-of-distribution detection and generates synthetic dialogs regarding the images via multimodal conditional text generation. GST then trains a dialog agent on the synthetic and the original VisDial data. As a result, GST scales the amount of training data up to an order of magnitude that of VisDial (1.2M $\rightarrow$ 12.9M QA data). For robust training of the synthetic dialogs, we also propose perplexity-based data selection and multimodal consistency regularization. Evaluation on VisDial v1.0 and v0.9 datasets shows that GST achieves new state-of-the-art results on both datasets. We further observe the robustness of GST against both visual and textual adversarial attacks. Finally, GST yields strong performance gains in the low-data regime. Code is available at https://github.com/gicheonkang/gst-visdial.

1. Introduction

Recently, there has been extensive research towards developing visually-grounded dialog systems [12, 13, 34, 36] due to their significance in many real-world applications (*e.g.,* helping visually impaired person). Notably, Visual Dialog (VisDial) [12] has provided a testbed for studying such systems, where a dialog agent should answer a *sequence* of image-grounded questions. For instance, the agent is expected to answer open-ended questions like *"What color is it?"* and *"How old does she look?"*. This task requires a holistic understanding of visual information, linguistic semantics in context (*e.g.,* it and she), and most importantly, the grounding of these two.

Most of the previous approaches in VisDial [9, 10, 18, 20, 25, 26, 30, 31, 35, 49, 54, 55, 64, 67, 78, 84] have trained the dialog agents solely on VisDial data via supervised learning. More recent studies [8,53,77] have employed self-supervised pre-trained models such as BERT [14] or ViLBERT [48] and finetuned them on VisDial data. The models are typically pre-trained to recover masked inputs and predict the semantic alignment between two segments. This *pretrain-then-transfer* learning strategy has shown promising results by transferring knowledge from the models pre-trained on large-scale data sources [4, 71, 85] to VisDial.

Our research question is the following: *How can the dialog agent expand its knowledge beyond what it can acquire via supervised learning or self-supervised pre-training on the provided datasets?* Some recent studies have shown that semi-supervised learning and pre-training have complementary modeling capabilities in image [86] and text classification [16]. Inspired by them, we consider semi-supervised learning (SSL) as a way to address the above question.

Let us assume that large amounts of unlabeled images are available. SSL for VisDial can be applied to generate synthetic conversations for the unlabeled images and train the agent with the synthetic data. However, there are two critical challenges to this approach. First, the target output for VisDial (*i.e.,* multi-turn visual QA data) is more complex than that of the aforementioned studies [16, 86]. Specifically, they have addressed the classification problems, yielding class probabilities as pseudo labels [39]. In contrast, SSL for VisDial should generate a sequence of pseudo queries (*i.e.,* visual questions) and pseudo labels (*i.e.,* corresponding answers) in *natural language* to train the answering agent. It further indicates that the target output should be generated while considering the *multimodal* and *sequential* nature of the visual dialog task. Next, even if SSL yields synthetic dialogs via text generation, there may be noise, such as generating irrelevant questions or incorrect answers to given contexts. A robust training method is required to leverage such noisy synthetic dialog datasets.

*Equal contribution

In this paper, we study the above challenges in the context of SSL, especially self-training [6, 16, 21, 28, 32, 39, 44, 52, 60, 65, 72, 73, 79, 80, 86], where a teacher model trained on labeled data predicts the pseudo labels for unlabeled data. Then, a student model jointly learns on the labeled and the pseudo-labeled datasets. Unlike existing studies in self-training that have mainly studied uni-modal, discriminative tasks such as image classification [72, 80, 86] or text classification [16, 32, 52], we extend the idea of self-training to the task of multimodal conditional text generation.

To this end, we propose a new learning strategy, called *Generative Self-Training* (GST), that artificially generates multi-turn visual QA data and utilizes the synthetic data for training. GST first trains the teacher model (answerer) and the visual question generation model (questioner) using VisDial data. It then retrieves a set of unlabeled images from a Web image dataset, Conceptual 12M [7]. Next, the questioner and the teacher generate a series of visual QA pairs for the retrieved images. Finally, the student is trained on the synthetic and the original VisDial data. We also propose perplexity-based data selection (PPL) and multimodal consistency regularization (MCR) to effectively train the student with the noisy dialog data. PPL is to selectively utilize the answers whose perplexity of the teacher is below a threshold. MCR encourages the student to yield consistent predictions when the perturbed multimodal inputs are given. As a result, GST successfully augments the synthetic VisDial data (11.7M QA pairs), thus mitigating the need to scale up the size of the human-annotated VisDial data, which is prohibitively expensive and time-consuming.

Our key contributions are three-fold. First, we propose Generative Self-Training (GST) that generates multi-turn visual QA data to leverage unlabeled Web images effectively. Second, experiments show that GST achieves new state-of-the-art performance on VisDial v1.0 and v0.9 datasets. We further demonstrate two important results: (1) GST is indeed effective when the human-annotated visual dialog data is extremely scarce (improving up to 11.09 absolute points on NDCG), and (2) PPL and MCR are effective when training the noisy synthetic dialog data. Third, to validate the robustness of GST, we evaluate our proposed method under three different visual and textual adversarial attacks, *i.e*., FGSM, coreference, and random token attacks. We observe that GST significantly improves the performance compared with the baseline models against all adversarial attacks, especially boosting NDCG scores from 21.60% to 45.43% in the FGSM attack [19].

2. Related work

Visual dialog. Visual Dialog (VisDial) [12] has been proposed as an extended version of Visual Question Answering (VQA) [3, 4, 33], where a dialog agent should answer a series of interdependent questions using an image and the dialog history. Prior work has developed a variety attention mechanisms [18, 20, 30, 35, 49, 54, 55, 64, 67, 78] considering the interactions among the image, dialog history, and question. Some studies [31, 84] have attempted to discover the semantic structures of the dialog in the context of graph neural networks [63] using the soft attention mechanisms [5]. From the learning algorithm perspective, all of them have relied on supervised learning on VisDial data. More recently, a line of research [8, 53, 77] has employed self-supervised pre-training to leverage the knowledge of related vision-and-language datasets [4, 71, 85]. However, our approach is based on semi-supervised learning and produces the task-specific data (*i.e.,* visual dialogs) for unlabeled images to train the dialog agent.

Sequence generation in vision-and-language tasks. Many studies have generated natural language for the visual inputs such as image captioning [3, 81], video captioning [23, 56], visual question generation (VQG) [17, 24, 29, 37, 47, 57], visual dialog (VisDial) [12, 18], and video dialog [2, 38]. Furthermore, recent studies [40, 82] have produced text data for vision-and-language pre-training. GST is similar to these studies in that the model generates the text data, but our focus is on studying the effect of semi-supervised learning (SSL) on top of such pre-training approaches. To the best of our knowledge, GST is the first approach to show the efficacy of SSL throughout a wide range of visual QA tasks.

Neural dialog generation. In NLP literature, extensive studies have been conducted regarding neural dialogue generation for both open-domain dialogue [41, 42, 62, 68, 70, 83] and task-oriented dialogue [22, 76]. Our approach is similar to neural dialogue generation in that the model should generate a corresponding response based on the dialog history and the current utterance. However, we aim to produce *visually-grounded* dialogs, and thus the image-groundedness of the question and the semantic correctness of the answer are important. On the other hand, neural dialogue generation considers many different aspects: specificity, response-relatedness [66], interestingness [50], and diversity [41].

3. Approach

3.1. Preliminaries

Self-training. We have a labeled dataset $L = \{(x_n, y_n)\}_{n=1}^{N}$ and an unlabeled dataset $U = \{\tilde{x}_m\}_{m=1}^{M}$. Typically, self-training trains a teacher model $P_{\mathcal{T}}$ on the labeled dataset L. The teacher then predicts the pseudo label $\tilde{y}$ for the unlabeled data $\tilde{x} \sim U$, constructing the pseudo-labeled dataset $\tilde{L} = \{(\tilde{x}_m, \tilde{y}_m)\}_{m=1}^{M}$. Finally, a student model $P_{\mathcal{S}}$ is trained on $L \cup \tilde{L}$. Many variants have been studied on this setup: (1) selecting the subset of the pseudo-labeled dataset [21, 72, 80], (2) adding noise to inputs [21, 72, 79, 80, 86], and (3) iterating the above setup multiple times [21, 80].

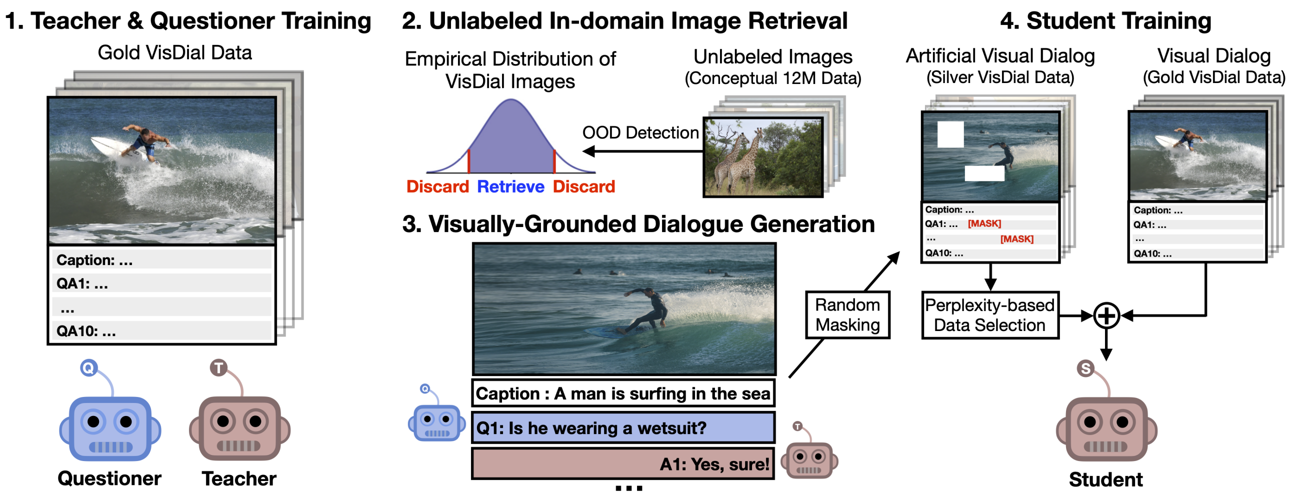


Figure 1. An overview of Generative Self-Training (GST).

Visual dialog. The visual dialog (VisDial) dataset [12] contains an image v and a visually-grounded dialog $d = \{\underbrace{c}_{d_0}, \underbrace{(q_1, a_1^{gt})}_{d_1}, \cdots, \underbrace{(q_T, a_T^{gt})}_{d_T}\}$ where c denotes an image caption. T is the number of rounds for each dialog. At round t, a dialog agent is given a triplet $(v, d_{<t}, q_t)$ as an input, consisting of the image, the dialog history, and a visual question. $d_{<t}$ denotes all dialog rounds before the t-th round. The agent is then expected to predict a ground-truth answer a_t^{gt}. There are two broad classes of methods in VisDial: *generative* and *discriminative*. Generative models aim to generate the ground-truth answer by maximizing the log-likelihood of a_t^{gt}. In contrast, discriminative models are trained to retrieve the ground-truth answer from a list of answer candidates $a_t^{gt} \in \{a_t^1, \cdots, a_t^{100}\}$. Our main focus is the generative models since they do not need pre-defined answer candidates and are thus more practical to be deployed in real-world applications.

3.2. Generative Self-Training (GST)

This subsection describes our approach, called GST, which generates multi-turn visual QA data and utilizes the generated data for training. An overview of GST is shown in Figure 1. We have a human-labeled VisDial dataset $L = \{(v_n, d_n)\}_{n=1}^N$ where v_n is a given image, and each dialog $d_n = \{\underbrace{c_n}_{d_{n,0}}, \underbrace{(q_{n,1}, a_{n,1}^{gt})}_{d_{n,1}}, \cdots, \underbrace{(q_{n,T}, a_{n,T}^{gt})}_{d_{n,T}}\}$ consists of an image caption c and T rounds of QA pairs. In the following, we omit the superscript gt in the ground-truth answer for brevity. GST first trains a teacher $P_\mathcal{T}$ and a questioner $P_\mathcal{Q}$ with the labeled dataset L via supervised learning. It then retrieves unlabeled images $U = \{\tilde{v}_m\}_{m=1}^M$ from the Conceptual 12M dataset [7] using a simple outlier detection model, the multivariate normal distribution. Next, the questioner and the teacher generate the visually-grounded dialog $\tilde{d}$ for the unlabeled image $\tilde{v}$ via multimodal conditional text generation, finally yielding a synthetic dialog dataset $\tilde{L} = \{(\tilde{v}_m, \tilde{d}_m)\}_{m=1}^M$. We call this dataset the *silver VisDial* data to distinguish it from the human-labeled VisDial dataset [12] (short for the *gold VisDial* data). Finally, a student $P_\mathcal{S}$ is trained on a combination of the gold and the silver VisDial data while applying perplexity-based data selection (PPL) and multimodal consistency regularization (MCR) to the silver VisDial data. We describe the details of each process in the following parts.

Teacher & questioner training. First, a series of question-and-answer pairs for the unlabeled images should be generated to train the answering agent. Accordingly, GST first trains the answer generator, the teacher model $P_\mathcal{T}$, on the gold VisDial dataset. Specifically, the teacher learns to generate the ground-truth answer's word sequence $a_t = (w_{t,1}, \cdots, w_{t,S})$, given the context triplet $c_t \triangleq (v, d_{<t}, q_t)$, consisting of the image, the dialog history, and the question. It is optimized by minimizing the negative log-likelihood of the ground-truth answer. Formally,

$$\begin{aligned} \mathcal{L}_\mathcal{T} &= -\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\log P_\mathcal{T}(a_{n,t}|c_{n,t}) \\ &= -\frac{1}{NTS}\sum_{n=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{S}\log P_\mathcal{T}(w_s|c_{n,t}, w_{<s}) \end{aligned} \tag{1}$$

where N, T, and S denote the number of data tuples in gold VisDial data, dialog rounds, and the sequence length of the ground-truth answer, respectively. $w_{<s}$ indicates all word tokens before the s-th token in the answer sequence. Similar to the teacher, the questioner is trained to generate the

question at round t, given the image and the dialog history until round $t-1$ (*i.e.,* $P_{\mathcal{Q}}(q_t|v, d_{<t})$). The questioner is also optimized by minimizing the negative log-likelihood of the follow-up question. Note that the teacher and the questioner are trained separately to prevent possible unintended co-adaptation [34]. Both the teacher and the questioner are based on encoder-decoder architecture, where an encoder aggregates the context triplet, and a decoder generates the target sentence. We implement the models by integrating a pre-trained vision-and-language encoder, ViLBERT [48], with the transformer decoder [61]. We refer readers to Appendix A for a detailed architecture.

Unlabeled in-domain image retrieval (IIR). Inspired by the work [16] that highlighted the importance of using in-domain data, GST retrieves in-domain image data from the Conceptual 12M dataset [7] with an out-of-distribution (OOD) detection model. Specifically, we extract the D dimensional feature vector for each image in the gold VisDial dataset by using the Vision Transformer (ViT) [15] in the CLIP model [58], yielding a feature matrix for the entire images $\mathbf{X} = (X_1, \cdots, X_N)^\top \in \mathbb{R}^{N \times D}$. Based on the matrix, we build the multivariate normal distribution whose dimension is D, *i.e.,* $\mathbf{X} \sim \mathcal{N}_D(\mu, \Sigma)$. We regard this normal distribution as the empirical distribution of the gold VisDial images and perform OOD detection by identifying the probability of each feature vector for the unlabeled image. Consequently, the top-M unlabeled images are retrieved out of 12 million Web images ($M \approx 3.6$ million).

Visually-grounded dialog generation. This step mimics a scenario where two people have a conversation about the given images. Given the retrieved images $U = \{\tilde{v}_m\}_{m=1}^M$, our goal is to generate the visually-grounded dialogs $\{\tilde{d}_m\}_{m=1}^M$ where each dialog $\tilde{d}$ consists of the image caption and T rounds of QA pairs. In an actual implementation, we use the image captions in the Conceptual 12M dataset [7] and thus do not generate the captions. The QA pairs are sequentially generated. Concretely, the image $\tilde{v}$, the caption $\tilde{c}$, and the generated QA pairs until round $t-1$ are used as inputs when the questioner generates the question at round t (*i.e.,* $\tilde{q}_t$). After then, the teacher produces the corresponding answer $\tilde{a}_t$ based on the image $\tilde{v}$, the dialog history $\tilde{d}_{<t}$, and the question $\tilde{q}_t$. Finally, GST produces the silver VisDial dataset $\tilde{L} = \{(\tilde{v}_m, \tilde{d}_m)\}_{m=1}^M$.

Student training with noisy data. In Figure 1, the student $P_{\mathcal{S}}$ is trained on the combination of the silver and the gold VisDial data. According to many studies [21, 72, 80, 86] in self-training, selectively utilizing the samples in the pseudo-labeled dataset is a common approach since the confidence of the teacher model's predictions varies from sample to sample. To this end, we introduce a simple yet effective data selection method for the sequence generation problem, perplexity-based data selection (PPL). PPL is to utilize the answers whose perplexity of the teacher is below a certain threshold. Perplexity is defined as the exponentiated average negative log-likelihood of a sequence; the lower, the better. We hypothesize that PPL, albeit noisy, can be an indicator of whether the generated answer is correct or not, as in [69]. Furthermore, inspired by the consistency regularization [72, 79] widely utilized in recent SSL algorithms, we also propose the multimodal consistency regularization (MCR) to improve the generalization capability of the student. MCR encourages the student to yield predictions similar to the teacher's predictions even when the student is provided with perturbed multimodal inputs. Finally, we design a loss function for the student as:

$$
\begin{aligned}
&\mathcal{L}_{\mathcal{S}} = \\
&-\frac{1}{MT}\sum_{m=1}^{M}\sum_{t=1}^{T}\mathbb{1}(\mathrm{PPL}(\tilde{a}_{m,t}) < \tau)\log \underbrace{P_{\mathcal{S}}(\tilde{a}_{m,t}|\mathcal{M}(\tilde{c}_{m,t}))}_{\mathrm{MCR}} \\
&-\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\log P_{\mathcal{S}}(a_{n,t}|c_{n,t}) \\
&\text{where } \mathrm{PPL}(\tilde{a}_t) = \exp\left\{-\frac{1}{S}\sum_{s=1}^{S}\log P_{\mathcal{T}}(\tilde{w}_s|\tilde{c}_t, \tilde{w}_{<s})\right\}
\end{aligned}
\tag{2}
$$

where M, $\mathbb{1}$, and τ denote the number of data tuples in silver VisDial data, indicator function, and selection threshold, respectively. $\tilde{c}_{m,t} \triangleq (\tilde{v}_m, \tilde{d}_{m,<t}, \tilde{q}_{m,t})$ denotes the context for the silver VisDial data. The loss function is the sum of the losses for the silver and the gold VisDial data. PPL and MCR are applied to compute the loss of the silver VisDial data. PPL is used in the indicator function above, selecting the synthetic answers whose perplexity of the teacher is below the threshold τ. It implies that the unselected answers are ignored during training. The teacher's perplexity of each answer is computed in the dialog generation step above. Next, $\mathcal{M}$ denotes the stochastic function for MCR that injects perturbations to the input space of the student. Inspired by ViLBERT [48], we implement the stochastic function by randomly masking 15% of image regions and word tokens. Specifically, masked image regions have their image features zeroed out, and the masked word tokens are replaced with a special `[MASK]` token. The intuition behind MCR is minimizing the distance between the *perturbed* (*i.e.,* masked) predictions from the student and the *unperturbed* predictions (*i.e.,* $\tilde{a}_{m,t}$) from the teacher. It indicates that the perturbation is not injected when the teacher generates the synthetic answers. We believe MCR makes the student robust to the input noise, and PPL encourages the student to maintain a low entropy (*i.e.,* confident) in noisy data training. The student and the teacher have the same model capacity and are based on the same model architecture.

4. Experiments

4.1. Experimental setup

VisDial datasets. We evaluate our proposed approach on the VisDial v1.0 and v0.9 datasets [12], collected by the AMT chatting between two workers about MS-COCO [46] images. Each dialog consists of a caption from COCO and a sequence of ten QA pairs. The VisDial v0.9 dataset has 83k dialogs on COCO-train and 40k dialogs on COCO-validation images. More recently, Das *et al.* [12] released additional 10k dialogs on Flickr images to use them as validation and test splits for the VisDial v1.0 dataset. As a result, the VisDial v1.0 dataset contains 123k, 2k, and 8k dialogs as train, validation, and test split. This dataset is licensed under a Creative Commons Attribution 4.0 International License.

Evaluation protocol. We follow the standard evaluation protocol established in the work [12] for evaluating visual dialog models. The visual dialog models for both generative and discriminative tasks have been evaluated by the retrieval-based evaluation metrics: mean reciprocal rank (MRR), recall@k (R@k), mean rank (Mean), and normalized discounted cumulative gain (NDCG). Specifically, all dialogs in VisDial contain a list of 100 answer candidates for each visual question, and there is one ground-truth answer in the answer candidates. The model sorts the answer candidates by the log-likelihood scores and then is evaluated by the four different metrics. MRR, R@k, and Mean consider the rank of the single ground-truth answer, while NDCG[1] considers all relevant answers from the 100-answers list by using the densely annotated relevance scores for all answer candidates. The community regards NDCG as the primary evaluation metric.

The size of synthetic data. The size of the silver VisDial data (*i.e.,* M) is 3.6M which is 30x larger than that of the gold VisDial data ($N = 0.12$M). Note that the silver VisDial data contains approximately 36M QA pairs since each dialog contains 10 QA pairs. 11.7M QA pairs out of 36M ($\sim$32%) are actually utilized after applying perplexity-based data selection when we set the selection threshold τ to 50. Consequently, the total amount of the training data is nearly 12.9M QA pairs, combining the silver data (11.7M QA pairs) with the original gold data (1.2M QA pairs).

Iterative training. We introduce the concept of iterative training [21, 80], which iterates the self-training algorithm a few times. The iterative training treats the student model at i-th iteration as a teacher model at (i+1)-th iteration to generate a new synthetic silver data and train a new student. Specifically, the iterative training repeats the third and fourth steps in Figure 1, where the silver VisDial data accumulates as the iteration proceeds. The student model at each iteration is trained with the accumulated silver and gold data by following the previous studies [21, 80]. We iterate GST up to three times. Unless stated otherwise, the student model is trained with three iterations.

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

4.2. Visual dialog results

Comparison with state-of-the-art. We compare GST with the state-of-the-art approaches on the validation set of the VisDial v1.0 and v0.9 datasets, consisting of UTC [8], MITVG [9], VD-BERT [77], LTMI [54], KBGN [25], DAM [26], ReDAN [18], DMRM [10], Primary [20], RvA [55], CorefNMN [35], CoAtt [78], HCIAE [49], and MN [12]. We decided to use the validation splits since all previous studies benchmarked the models on those splits. In Table 1, GST significantly outperforms all compared methods on all evaluation metrics. Compared with the state-of-the-art model, the student model improves MRR 3.20% (56.83 $\rightarrow$ 60.03) and R@1 3.26% (47.14 $\rightarrow$ 50.40) on the VisDial v0.9 dataset. The improvement is consistently observed on the VisDial v1.0 dataset, boosting NDCG 1.61% (63.86 $\rightarrow$ 65.47) and MRR 0.97% (52.22 $\rightarrow$ 53.19). Moreover, it is noticeable that recent strong models (*i.e.,* UTC, MITVG, and VD-BERT) are also built based on the pre-trained weights of ViLBERT [48], transformer [75], and BERT [14], respectively. Our proposed method also achieves new state-of-the-art results on the discriminative VisDial models. Details can be found in Appendix B.

GST in the low-data regime. Is GST also helpful when gold data is scarce? We investigate this question to identify the effect of GST in the low-data regime. We assume that only a small subset of the gold VisDial data (1%, 5%, 10%, 20%, and 30%) is available. Therefore, the size of the gold data is 0.01N, 0.05N, 0.1N, 0.2N, and 0.3N, respectively. We first train the teacher and the questioner on such scarce data, and then these two agents generate a new silver VisDial data for unlabeled images in the Conceptual 12M dataset [7] with size $5N$. The student is then trained on the newly generated silver VisDial data and the small amount of the gold VisDial data. The student is based on a single iterative training, and PPL and MCR are still applied in this experiment. In Table 2, GST yields huge improvements on both metrics, especially NDCG, boosting up to 11.09 absolute points compared with the teacher. We observe that the smaller the amount of gold data, the larger the performance gap between the teacher and the student on NDCG. It implies that GST is helpful, especially when gold data is scarce. We speculate the results in the low-data regime are particularly remarkable in other dialog-based tasks [2, 45, 59, 74] since they are based on relatively small-scaled datasets, and scaling up the size of the human-dialog datasets is laborious and expensive.

Model	VisDial v0.9 (val)					VisDial v1.0 (val)					
	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓	NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
MN† [12]	52.59	42.29	62.85	68.88	17.06	51.86	47.99	38.18	57.54	64.32	18.60
HCIAE† [49]	53.86	44.06	63.55	69.24	16.01	59.70	49.07	39.72	58.23	64.73	18.43
CoAtt† [78]	55.78	46.10	65.69	71.74	14.43	59.24	49.64	40.09	59.37	65.92	17.86
CorefNMN [35]	53.50	43.66	63.54	69.93	15.69	-	-	-	-	-	-
RvA [55]	55.43	45.37	65.27	72.97	**10.71**	-	-	-	-	-	-
Primary [20]	-	-	-	-	-	-	49.01	38.54	59.82	66.94	16.60
DMRM [10]	55.96	46.20	66.02	72.43	13.15	-	50.16	40.15	60.02	67.21	15.19
ReDAN [18]	-	-	-	-	-	60.47	50.02	40.27	59.93	66.78	17.40
DAM [26]	-	-	-	-	-	60.93	50.51	40.53	60.84	67.94	16.65
KBGN [25]	-	-	-	-	-	60.42	50.05	40.40	60.11	66.82	17.54
LTMI [54]	-	-	-	-	-	63.58	50.74	40.44	61.61	69.71	14.93
VD-BERT [77]	55.95	46.83	65.43	72.05	13.18	-	-	-	-	-	-
MITVG [9]	<u>56.83</u>	<u>47.14</u>	<u>67.19</u>	<u>73.72</u>	<u>11.95</u>	61.47	51.14	41.03	61.25	68.49	<u>14.37</u>
UTC [8]	-	-	-	-	-	<u>63.86</u>	<u>52.22</u>	<u>42.56</u>	<u>62.40</u>	<u>69.51</u>	15.67
Student (ours)	**60.03**±.18	**50.40**±.15	**70.74**±.09	**77.15**±.13	12.13±.18	**65.47**±.14	**53.19**±.11	**43.08**±.10	**64.09**±.05	**71.51**±.13	**14.34**±.15

Table 1. Comparison with the state-of-the-art generative models on both VisDial v0.9 and v1.0 validation datasets. ↑ indicates higher is better. ↓ indicates lower is better. NDCG is not supported in v0.9 dataset. † denotes that the models are re-implemented by the previous work [18]. The standard deviations of our proposed models are reported ± with three different initialized models.

Model	NDCG				
	1%	5%	10%	20%	30%
Teacher	27.64	50.04	54.46	57.14	60.67
Student	**38.73** (**+11.09**)	**56.60** (**+6.56**)	**58.62** (**+4.16**)	**60.92** (**+3.78**)	**63.09** (**+2.42**)

Table 2. Results of GST in the low-data regime. We report NDCG scores based on the VisDial v1.0 validation split. We assume a small subset of the gold VisDial data (∼30%) is available.

Model	Question Type					
	Yes / No (60.4%)	Color (14.8%)	Objects (5.1%)	Counting (3.1%)	Time / Place (8.5%)	Others (9.0%)
Teacher	66.87	60.61	53.67	49.44	69.36	61.32
Student	**67.41** (**+0.54**)	**61.85** (**+1.24**)	**55.25** (**+1.58**)	**51.76** (**+2.32**)	**71.38** (**+2.02**)	**63.02** (**+1.70**)

Table 3. Question type analysis on the VisDial v1.0 validation split. The percentage denotes the data proportion of each category.

Question type analysis. We conduct a question-type analysis to identify what type of questions obtain benefits from GST. We divided the question type into six categories, *Yes/No, Color, Objects, Counting, Time/Place,* and *Others*. In Table 3, the student model obtains more gains compared with the teacher model in less frequent question types (*e.g.,* Counting and Time / Place).

4.3. Adversarial robustness results

We introduce a comprehensive evaluation setup for adversarial robustness in VisDial. Specifically, we propose three different adversarial attacks: (1) the FGSM attack, (2) a coreference attack, and (3) a random token attack. The FGSM attack perturbs input visual features, and the others attack the dialog history (*i.e.,* textual inputs).

Baselines. We compare our student model against two ablative baselines: (1) the teacher model and (2) the student model utilizing the entire CC12M images without applying the in-domain image retrieval (*i.e.,* student-iter1-full). We propose the student-iter1-full model to study the effect of the discarded images and the corresponding synthetic dialog data on adversarial robustness.

Adversarial robustness against the FGSM attack. The Fast Gradient Signed Method (FGSM) [19] is a white-box attack that perturbs the visual inputs based on the gradients of the loss with respect to the visual inputs. Formally,

$$\mathrm{FGSM}(x) = x + \epsilon \cdot \mathrm{sign}(\nabla_x \mathcal{L}(x, y)) \tag{3}$$

where x and y denote the visual inputs and the corresponding ground-truth labels, respectively. ϵ is a hyperparameter that adjusts the intensity of perturbations. However, different from the above setup, each question in VisDial can have one or more relevant answers in the list of answer candidates. We thus define the FGSM attack for VisDial as follows:

$$\mathrm{FGSM}(v) = v + \epsilon \cdot \mathrm{sign}(\sum_{c=1}^{C} r(a_{t,c}) \cdot \nabla_v \mathcal{L}(c_t, a_{t,c})) \tag{4}$$

where $C = 100$ and $r(\cdot)$ denote the number of answer candidates and a function that returns the human-annotated relevance scores for each answer candidate, respectively. The relevance scores range from 0 to 1. c_t and $a_{t,c}$ are the context triplet (*i.e.,* $c_t \triangleq (v, d_{<t}, q_t)$) and the c-th answer candidate, respectively. The Equation 4 indicates that the gradients of the loss for all relevant answers are considered for the FGSM attack.

Model	No Attack	Coreference Attack	Random Token Attack			
			10%	20%	30%	40%
Teacher	56.55	52.60	54.69±1.12	52.86±0.79	49.41±2.09	45.04±2.28
Student (iter1, full)	58.53	54.26	56.59±1.37	54.55±1.15	50.98±2.06	46.56±1.96
Student (iter1)	58.63	54.34	55.59±0.88	54.26±1.54	51.04±2.39	47.04±2.03
Student (iter2)	56.92	52.69	55.59±0.88	53.57±1.40	49.95±1.91	46.82±2.02
Student (iter3)	**59.30**	**55.44**	**57.25**±0.91	**55.10**±1.50	**52.11**±2.75	**48.00**±2.90

Table 4. Adversarial robustness results against the attacks on the dialog history. We apply two different dialog history attacks: a coreference attack and a random token attack. The models are evaluated on the VisDialConv dataset [1] with the NDCG metric. The standard deviations are reported ± with five different random seeds.

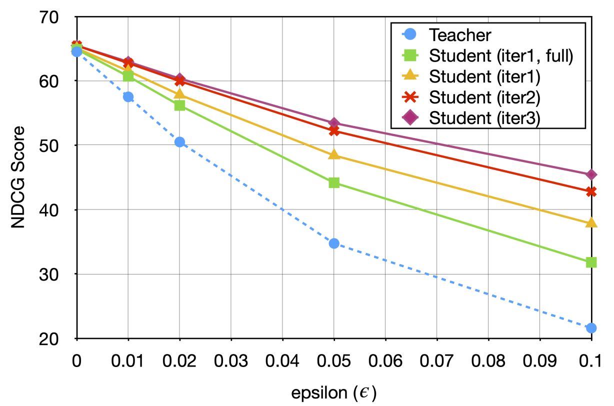


Figure 2. Adversarial robustness against FGSM attack on VisDial v1.0 validation split. We report NDCG scores of each model.

As shown in Figure 2, we validate the models with four different epsilon values $\epsilon \in \{0.01, 0.02, 0.05, 0.1\}$. The student model shows very significant improvements in NDCG compared with the teacher model. Specifically, the performance gap between the student model with three iterations (*i.e.,* student-iter3) and the teacher model widens up to 23.83 absolute points (21.60 $\rightarrow$ 45.43) when ϵ is 0.1. It illustrates that GST makes the visual dialog model robust against the FGSM attack even though the student model is not optimized for adversarial robustness. Furthermore, we can clearly identify the efficacy of the iterative training as the intensity of the perturbations increases. The NDCG scores are boosted from 37.82% (iter1) to 45.43% (iter3) at $\epsilon = 0.1$. Finally, the student-iter1 model shows better performance than the student-iter1-full model. It implies that the additional use of the discarded images along with the synthetic dialog does not bring any gains in the FGSM attack.

Adversarial robustness against the textual attacks. We also study the adversarial robustness against textual attacks to illustrate the effect of GST. We decide to perturb the dialog history because it contains useful information to answer the given question (*e.g.,* cues for pronoun). However, according to recent studies [1, 31] in VisDial, not all questions require the dialog history to respond with the correct answers. So the work [1] has proposed a challenging subset of the VisDial validation dataset called VisDialConv. The VisDialConv dataset only contains questions that necessarily require the dialog history to answer (*e.g.,* can you tell what it is for?). The crowd-workers conducted a manual inspection to select such *context-dependent* questions.

Based on the VisDialConv dataset, we apply two different black-box attacks. First, we propose the coreference attack, which substitutes the noun phrases or pronouns in the dialog history with their synonyms to fool the VisDial models. Specifically, we leverage the off-the-shelf neural coreference resolution tool[2] and find words in the dialog history that refer to objects such as those mentioned in a given question. We also borrow the counter-fitting word embeddings [51] similar to textfooler [27] to retrieve the synonyms. We greedily substitute the words with the synonyms with a minimum cosine distance in the embedding space since we observe that the other synonyms harm the original semantics of the dialog history. In Table 4, the student-iter3 model outperforms the teacher model on NDCG by a large margin (2.84%, 52.60 $\rightarrow$ 55.44) in the coreference attack. Furthermore, we do not see any merit in utilizing the entire CC12M [7] images and the corresponding synthetic dialog data, comparing the student-iter1-full with the student-iter1.

The random token attack randomly replaces the word or sub-word tokens in the dialog history with a special `[MASK]` token. The pre-trained BERT$_{\texttt{BASE}}$ model [14] then recovers the masked tokens with masked language modeling (MLM) similar to BERT-ATTACK [43]. Finally, the perturbed dialog history is fed into the visual dialog models. We conduct this experiment by adjusting the probability of random masking up to 40%. As shown in Table 4, we evaluate each model with five random seeds and report the arithmetic mean and the standard deviations. The results demonstrate that GST is relatively robust against the random token attack compared with the baseline models.

[2] https://github.com/huggingface/neuralcoref based on the work [11].

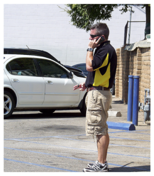
Image & Caption	Human Conversation (Gold VisDial)	Machine Conversation (Silver VisDial)
 A man is talking on a cell phone in a parking lot	D1 : Is it daytime? Yes D2 : Is this is a busy full parking lot? No D3 : Is the man dressed for work? No D4 : What is he wearing? Polo shirt and shorts D5 : Are there other people around? No D6 : Do you see any cars? Yes D7 : Are there a lot of cars? No D8 : How many cars do you think there are? 3 D9 : What colors are the cars? White and black D10 : Is the man wearing sunglasses? Yes	D1 : Is the man wearing a hat? No D2 : Is the man dressed for cold or warm weather? Warm weather D3 : Is the man talking on a cell phone? Yes he is D4 : Is it a flip phone? Yes it is D5 : What color is the phone? Black and silver D6 : Is the man holding it in his hand? Yes it is D7 : Are there any cars in the parking lot? Yes D8 : Is he waiting for someone to get on the car? No just standing D9 : Is the man looking at the camera? No he is looking off to the side D10 : Are there any buildings in the background? \ There is a brick building behind him
 A woman rides her bike on a sidewalk next to a fence	D1 : Is it daytime? Yes D2 : Is the sun out? Yes D3 : How old is the woman? 24 D4 : Is she alone? Yes D5 : Is she wearing sunglasses? No D6 : Is she carrying anything else with her? No D7 : Is she wearing sneakers? Yes D8 : What color is her shirt? Wearing a black jacket D9 : Is she smiling? No D10 : Where is she going? West on the road	D1 : Is this a color photo? Yes, it is a color photo D2 : Is it daytime? Yes D3 : What is the weather like? Looks like sunny D4 : What color is the woman's hair? Hair is brown D5 : How old does she seem? Around 35 D6 : Is she wearing a helmet? No helmet seen D7 : Is she wearing jeans? No D8 : Is she wearing glasses? No D9 : What color is her bike? Bike is blue in color D10 : Are there any cars? Yes, i see 1 car

Figure 3. A visualization of the gold and the silver data on VisDial v1.0 validation split.

4.4. Analysis of the silver VisDial data

Comparison between silver and gold data. For qualitative analysis of the silver data, we visualize the generated conversations from our proposed models and the ones from humans. We excerpt the human conversation from the VisDial v1.0 validation dataset, and the questioner and the student generate the machine conversation using the image and the caption in the validation data. As shown in Figure 3, diverse visual questions are generated in the silver VisDial data. For example, in D10 of the last example, the questioner asks about "a car" not mentioned by the human questioner and not even presented in the image caption. The student responds correctly to the question. Likewise, from D3 to D6 in the first example, the questioner deals with "a cell phone," whereas the human questioner deals with different topics. However, we identify that the student sometimes fails to generate correct answers (*i.e.,* the red-colored text), showing the importance of more precise visual grounding.

The diversity of silver questions. We further quantify the generated question's diversity by comparing the gold questions with the silver ones for the same images in the VisDial v1.0 validation dataset. We extract N-grams for every ten questions (*i.e.,* per image) in the gold and silver data and compare the N-grams between the two. We define the question diversity as the percentage of *unique* silver N-grams not observed in the gold N-grams. We identify the question diversity by adjusting N from one to four. We generate three silver datasets and report the mean and standard deviations of the question diversity since the questioner performs stochastic decoding (see Appendix D). In Table 5, the diversity significantly increases as N increases (92.80% at N=4). It indicates that the questioner mainly generates different and distinctive 4-grams compared with the human questioner. Furthermore, as shown in No Match at Table 5, the questioner rarely generates the same questions that belong to gold questions. We analyze the answer diversity in Appendix C.

Model	N-gram Diversity				No Match
	N=1	N=2	N=3	N=4	
Questioner	**28.06** $\pm$0.14	**56.46** $\pm$0.09	**76.98** $\pm$0.08	**92.80** $\pm$0.08	**95.38** $\pm$0.15

Table 5. The N-gram diversity of the generated questions on the VisDial v1.0 validation images. The standard deviations are reported $\pm$ with three silver datasets. No match denotes the percentage of silver questions that do not precisely match the gold questions.

4.5. Ablation study

The results of an ablation study are in Appendix B.2.

5. Conclusion

We propose a semi-supervised learning approach for VisDial, called GST, that generates a synthetic visual dialog dataset for unlabeled Web images via multimodal conditional text generation. GST achieves the new state-of-the-art performance on the VisDial v1.0 and v0.9 datasets. Moreover, we demonstrate the efficacy of GST in low-data regime and adversarial robustness analysis. Finally, GST produces diverse dialogs compared with the human dialog. We believe the idea of GST is generally applicable to other multimodal generative domains and expect GST to open the door to leveraging unlabeled images in many visual QA tasks.

Acknowledgements. This work was supported by the SNU-NAVER Hyperscale AI Center and the Institute of Information & Communications Technology Planning & Evaluation (IITP) (2021-0-01343-GSAI/40%, 2022-0-00953-PICA/30%, 2022-0-00951-LBA/20%, 2021-0-02068-AIHub/10%) grant funded by the Korean government.

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. 7

[2] Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. Audio visual scene-aware dialog. In *CVPR*, 2019. 2, 5

[3] 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. 2

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 1, 2

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2014. 2

[6] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*, 2019. 2

[7] 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. 2, 3, 4, 5, 7

[8] 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. 1, 2, 5, 6

[9] Feilong Chen, Fandong Meng, Xiuyi Chen, Peng Li, and Jie Zhou. Multimodal incremental transformer with visual grounding for visual dialogue generation. In *ACL*, 2021. 1, 5, 6

[10] Feilong Chen, Fandong Meng, Jiaming Xu, Peng Li, Bo Xu, and Jie Zhou. Dmrm: A dual-channel multi-hop reasoning model for visual dialog. In *AAAI*, 2020. 1, 5, 6

[11] Kevin Clark and Christopher D Manning. Deep reinforcement learning for mention-ranking coreference models. In *ACL*, 2016. 7

[12] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *CVPR*, 2017. 1, 2, 3, 5, 6

[13] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*, 2017. 1

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 1, 5, 7

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 4

[16] Jingfei Du, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Ves Stoyanov, and Alexis Conneau. Self-training improves pre-training for natural language understanding. In *NAACL*, 2021. 1, 2, 4

[17] Zhihao Fan, Zhongyu Wei, Piji Li, Yanyan Lan, and Xuanjing Huang. A question type driven framework to diversify visual question generation. In *IJCAI*, 2018. 2

[18] Zhe Gan, Yu Cheng, Ahmed El Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. Multi-step reasoning via recurrent dual attention for visual dialog. In *ACL*, 2019. 1, 2, 5, 6

[19] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 2, 6

[20] Dalu Guo, Chang Xu, and Dacheng Tao. Image-question-answer synergistic network for visual dialog. In *CVPR*, 2019. 1, 2, 5, 6

[21] Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. Revisiting self-training for neural sequence generation. In *ICLR*, 2020. 2, 4, 5

[22] Xinting Huang, Jianzhong Qi, Yu Sun, and Rui Zhang. Mala: Cross-domain dialogue generation with action learning. In *AAAI*, 2020. 2

[23] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *CVPR Workshops*, 2020. 2

[24] Unnat Jain, Ziyu Zhang, and Alexander G Schwing. Creativity: Generating diverse questions using variational autoencoders. In *CVPR*, 2017. 2

[25] Xiaoze Jiang, Siyi Du, Zengchang Qin, Yajing Sun, and Jing Yu. Kbgn: Knowledge-bridge graph network for adaptive vision-text reasoning in visual dialogue. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1265–1273, 2020. 1, 5, 6

[26] Xiaoze Jiang, Jing Yu, Yajing Sun, Zengchang Qin, Zihao Zhu, Yue Hu, and Qi Wu. Dam: Deliberation, abandon and memory networks for generating detailed and non-repetitive responses in visual dialogue. In *IJCAI*, 2020. 1, 5, 6

[27] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *AAAI*, 2020. 7

[28] Hwiyeol Jo and Ceyda Cinarel. Delta-training: Simple semi-supervised text classification using pretrained word embeddings. In *EMNLP*, 2019. 2

[29] Shen Kai, Lingfei Wu, Siliang Tang, Yueting Zhuang, Zhuoye Ding, Yun Xiao, Bo Long, et al. Learning to generate visual questions with noisy supervision. In *NeurIPS*, 2021. 2

[30] Gi-Cheon Kang, Jaeseo Lim, and Byoung-Tak Zhang. Dual attention networks for visual reference resolution in visual dialog. In *EMNLP*, 2019. 1, 2

[31] 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. 1, 2, 7

[32] Giannis Karamanolakis, Subhabrata Mukherjee, Guoqing Zheng, and Ahmed Hassan Awadallah. Self-training with weak supervision. In *NAACL*, 2021. 2

[33] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *NeurIPS*, volume 31, 2018. 2

[34] Jin-Hwa Kim, Nikita Kitaev, Xinlei Chen, Marcus Rohrbach, Byoung-Tak Zhang, Yuandong Tian, Dhruv Batra, and Devi Parikh. Codraw: Collaborative drawing as a testbed for grounded goal-driven communication. In *ACL*, 2019. 1, 4

[35] 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. 1, 2, 5, 6

[36] Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Clevr-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. In *NAACL*, 2019. 1

[37] Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Information maximizing visual question generation. In *CVPR*, 2019. 2

[38] Hung Le, Doyen Sahoo, Nancy F Chen, and Steven CH Hoi. Multimodal transformer networks for end-to-end video-grounded dialogue systems. In *ACL*, 2019. 2

[39] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on challenges in representation learning*, 2013. 1, 2

[40] 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. 2

[41] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. In *EMNLP*, 2016. 2

[42] Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. Adversarial learning for neural dialogue generation. In *EMNLP*, 2017. 2

[43] Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. Bert-attack: Adversarial attack against bert using bert. In *EMNLP*, 2020. 7

[44] Xinzhe Li, Qianru Sun, Yaoyao Liu, Qin Zhou, Shibao Zheng, Tat-Seng Chua, and Bernt Schiele. Learning to self-train for semi-supervised few-shot classification. In *NeurIPS*, volume 32, 2019. 2

[45] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. In *IJCNLP*, 2017. 5

[46] 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. 5

[47] Feng Liu, Tao Xiang, Timothy M Hospedales, Wankou Yang, and Changyin Sun. ivqa: Inverse visual question answering. In *CVPR*, 2018. 2

[48] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 1, 4, 5

[49] Jiasen Lu, Anitha Kannan, Jianwei Yang, Devi Parikh, and Dhruv Batra. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. In *NIPS*, 2017. 1, 2, 5, 6

[50] Shikib Mehri and Maxine Eskenazi. Unsupervised evaluation of interactive dialog with dialogpt. In *SIGDIAL*, 2020. 2

[51] Nikola Mrkšić, Diarmuid O Séaghdha, Blaise Thomson, Milica Gašić, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. Counter-fitting word vectors to linguistic constraints. In *NAACL*, 2016. 7

[52] Subhabrata Mukherjee and Ahmed Hassan Awadallah. Uncertainty-aware self-training for text classification with few labels. In *NeurIPS*, 2020. 2

[53] 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. 1, 2

[54] Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. Efficient attention mechanism for visual dialog that can handle all the interactions between multiple inputs. In *ECCV*, 2020. 1, 2, 5, 6

[55] Yulei Niu, Hanwang Zhang, Manli Zhang, Jianhong Zhang, Zhiwu Lu, and Ji-Rong Wen. Recursive visual attention in visual dialog. In *CVPR*, 2019. 1, 2, 5, 6

[56] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. Video captioning with transferred semantic attributes. In *CVPR*, 2017. 2

[57] Badri N Patro, Sandeep Kumar, Vinod K Kurmi, and Vinay P Namboodiri. Multimodal differential network for visual question generation. In *EMNLP*, 2018. 2

[58] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4

[59] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *ACL*, 2019. 5

[60] Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. Semi-supervised self-training of object detection models. In *IEEE Workshops on Application of Computer Vision*, 2005. 2

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

[62] Abdelrhman Saleh, Natasha Jaques, Asma Ghandeharioun, Judy Shen, and Rosalind Picard. Hierarchical reinforcement learning for open-domain dialog. In *AAAI*, 2020. 2

[63] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. In *IEEE Transactions on Neural Networks*. IEEE, 2008. 2

[64] Idan Schwartz, Seunghak Yu, Tamir Hazan, and Alexander G Schwing. Factor graph attention. In *CVPR*, 2019. 1, 2

[65] Henry Scudder. Probability of error of some adaptive pattern-recognition machines. In *IEEE Transactions on Information Theory*, 1965. 2

[66] Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. What makes a good conversation? how controllable attributes affect human judgments. In *NAACL*, 2019. 2

[67] Paul Hongsuck Seo, Andreas Lehrmann, Bohyung Han, and Leonid Sigal. Visual reference resolution using attention memory for visual dialog. In *NIPS*, 2017. 1, 2

[68] 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. 2

[69] Siamak Shakeri, Cicero Nogueira dos Santos, Henry Zhu, Patrick Ng, Feng Nan, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. End-to-end synthetic data generation for domain adaptation of question answering systems. In *EMNLP*, 2020. 4

[70] Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. In *ACL*, 2015. 2

[71] 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. 1, 2

[72] 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. 2, 4

[73] Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *NAACL*, 2021. 2

[74] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. In *CoRL*, 2020. 5

[75] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 5

[76] Kai Wang, Junfeng Tian, Rui Wang, Xiaojun Quan, and Jianxing Yu. Multi-domain dialogue acts and response co-generation. In *ACL*, 2020. 2

[77] 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. 1, 2, 5, 6

[78] Qi Wu, Peng Wang, Chunhua Shen, Ian Reid, and Anton Van Den Hengel. Are you talking to me? reasoned visual dialog generation through adversarial learning. In *CVPR*, 2018. 1, 2, 5, 6

[79] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. In *NeurIPS*, 2020. 2, 4

[80] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *CVPR*, 2020. 2, 4, 5

[81] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 2

[82] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *ICCV*, 2021. 2

[83] 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. 2

[84] Zilong Zheng, Wenguan Wang, Siyuan Qi, and Song-Chun Zhu. Reasoning visual dialogs with structural and partial observations. In *CVPR*, 2019. 1, 2

[85] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*, 2015. 1, 2

[86] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. In *NeurIPS*, 2020. 1, 2, 4