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

Gi-Cheon Kang\textsuperscript{1,2} Sungdong Kim\textsuperscript{3*} Jin-Hwa Kim\textsuperscript{3,2*} Donghyun Kwak\textsuperscript{4*} Byoung-Tak Zhang\textsuperscript{1,2} \\
\textsuperscript{1}IPAI, Seoul National University \quad \textsuperscript{2}AIIS \quad \textsuperscript{3}NAVER AI Lab \quad \textsuperscript{4}NAVER Cloud CLOVA \\
{chonkang, btzhang}@snu.ac.kr \quad \{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 → 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 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 [32, 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 $U = \{x_m\}_{m=1}^M$. Typically, self-training trains a teacher model $P_T$ on the labeled dataset $L$. The teacher then predicts the pseudo label $\hat{y}$ for the unlabeled data $x \sim U$, constructing the pseudo-labeled dataset $\hat{L} = \{(x_m, \hat{y}_m)\}_{m=1}^M$. Finally, a student model $P_S$ is trained on $L \cup \hat{L}$. 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].
Visual dialog. The visual dialog (VisDial) dataset [12] contains an image \( v \) and a visually-grounded dialog \( d = \{ c, (q_1, a_{gt}^1), \ldots, (q_T, a_{gt}^T) \} \) 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 \( a_{gt}^t \). 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 \( a_{gt}^t \). In contrast, discriminative models are trained to retrieve the ground-truth answer from a list of answer candidates \( a_{gt}^t \in \{a_1, \ldots, a_{100} \} \). 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 \( d_n = \{ c_n, (q_{n,1}, a_{gt,n,1}^1), \ldots, (q_{n,T}, a_{gt,n,T}^T) \} \) 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 \( P_T \) and a questioner \( P_Q \) 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 \( P_S \) 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 \( P_T \), on the gold VisDial dataset. Specifically, the teacher learns to generate the ground-truth answer’s word sequence \( a_t = (w_{t,1}, \ldots, w_{t,S}) \), given the context triplet \( c_t = (v, d_{<t}, q_t) \), 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,

\[
L_T = -\frac{1}{NT} \sum_{n=1}^N \sum_{t=1}^T \log P_T (a_{n,t} | c_{n,t})
\]

\[
= -\frac{1}{NTS} \sum_{n=1}^N \sum_{t=1}^T \sum_{s=1}^S \log P_T (w_s | c_{n,t}, w_{<s})
\]

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., $P_Q(q_t | v, d_{<t})$). 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 usefulness 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 $X = (X_1, \ldots, X_N)^\top \in \mathbb{R}^{N \times D}$. Based on the matrix, we build the multivariate normal distribution whose dimension is $D$, i.e., $X \sim \mathcal{N}(\mu, \Sigma)$. 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 that, 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 $P_S$ 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:

$$L_S = - \frac{1}{MT} \sum_{m=1}^M \sum_{t=1}^T \log P_S(\tilde{a}_{m,t} | \mathcal{M}(\tilde{c}_{m,t})) \quad \text{MCR}$$

$$- \frac{1}{NT} \sum_{n=1}^N \sum_{t=1}^T \log P_S(\tilde{a}_{n,t} | \mathcal{C}_n)$$

where $\text{PPL}(\tilde{a}_t) = \exp \left\{ \frac{1}{S} \sum_{s=1}^S \log P_T(\tilde{v}_s, \tilde{c}_t, \tilde{w}_{<s}) \right\}$

where $M$, $I$, and $\tau$ denote the number of data tuples in silver VisDial data, indicator function, and selection threshold, respectively. $\mathcal{C}_n \triangleq (\tilde{v}_m, \tilde{d}_{m,<t}, \tilde{q}_m)$ 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 $\tau$. 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 NDCG1 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 ($N = 0.12M$). 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 $\tau$ 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.

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], LTM [54], KBGN [25], DAM [26], ReDAN [18], DMRM [10], Primary [20], RvA [55], CorefNMN [35], CoAtt [78], HCI-AE [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 → 60.03) and R@1 3.26% (47.14 → 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 → 65.47) and MRR 0.97% (52.22 → 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.

---

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

<table>
<thead>
<tr>
<th>Model</th>
<th>VisDial v0.9 (val)</th>
<th>VisDial v1.0 (val)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MRR†</td>
<td>R@1†</td>
</tr>
<tr>
<td>MN [12]</td>
<td>52.59</td>
<td>42.29</td>
</tr>
<tr>
<td>HCIAE [49]</td>
<td>53.86</td>
<td>40.16</td>
</tr>
<tr>
<td>CoAtt [78]</td>
<td>55.78</td>
<td>46.10</td>
</tr>
<tr>
<td>CorefNMN [35]</td>
<td>53.50</td>
<td>43.66</td>
</tr>
<tr>
<td>RvA [55]</td>
<td>55.43</td>
<td>45.37</td>
</tr>
<tr>
<td>Primary [20]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DMRM [10]</td>
<td>55.96</td>
<td>46.20</td>
</tr>
<tr>
<td>ReDAN [18]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DAM [26]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>KBGN [25]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LTI [54]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>VD-BERT [77]</td>
<td>55.95</td>
<td>46.83</td>
</tr>
<tr>
<td>MITVG [9]</td>
<td>56.83</td>
<td>47.14</td>
</tr>
<tr>
<td>UTC [8]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Student (ours)</td>
<td>60.03±1.14</td>
<td>50.40±1.15</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Model</th>
<th>NDCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher</td>
<td>27.64</td>
</tr>
<tr>
<td></td>
<td>50.04</td>
</tr>
<tr>
<td></td>
<td>54.46</td>
</tr>
<tr>
<td></td>
<td>57.14</td>
</tr>
<tr>
<td></td>
<td>60.67</td>
</tr>
<tr>
<td>Student</td>
<td>38.73</td>
</tr>
<tr>
<td></td>
<td>56.60</td>
</tr>
<tr>
<td></td>
<td>58.62</td>
</tr>
<tr>
<td></td>
<td>60.92</td>
</tr>
<tr>
<td></td>
<td>63.09</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Question Type</th>
<th>Yes / No</th>
<th>Color</th>
<th>Objects</th>
<th>Counting</th>
<th>Time / Place</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher</td>
<td>66.87</td>
<td>60.61</td>
<td>53.67</td>
<td>49.44</td>
<td>69.36</td>
<td>61.32</td>
</tr>
<tr>
<td>Student</td>
<td>67.41</td>
<td>61.85</td>
<td>55.25</td>
<td>51.76</td>
<td>71.38</td>
<td>63.02</td>
</tr>
</tbody>
</table>

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,

\[ \text{FGSM}(x) = x + \epsilon \cdot \text{sign}(\nabla_x L(x, y)) \]  

where \( x \) and \( y \) denote the visual inputs and the corresponding ground-truth labels, respectively. \( \epsilon \) 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:

\[ \text{FGSM}(v) = v + \epsilon \cdot \text{sign} \left( \sum_{c=1}^{C} r(a_{t,c}) \cdot \nabla_v L(c_t, a_{t,c}) \right) \]  

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., context triplet \( (c_t, a_{t,c}) \)) 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.
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.

<table>
<thead>
<tr>
<th>Model</th>
<th>No Attack</th>
<th>Coreference Attack</th>
<th>Random Token Attack</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10%</td>
<td>20%</td>
<td>30%</td>
</tr>
<tr>
<td>Teacher</td>
<td>56.55</td>
<td>52.60</td>
<td>54.69 ± 1.12</td>
</tr>
<tr>
<td>Student (iter1, full)</td>
<td>58.53</td>
<td>54.26</td>
<td>56.59 ± 1.37</td>
</tr>
<tr>
<td>Student (iter1)</td>
<td>58.63</td>
<td>54.34</td>
<td>55.59 ± 0.88</td>
</tr>
<tr>
<td>Student (iter2)</td>
<td>56.92</td>
<td>52.69</td>
<td>55.59 ± 0.88</td>
</tr>
<tr>
<td>Student (iter3)</td>
<td>59.30</td>
<td>55.44</td>
<td>57.25 ± 0.91</td>
</tr>
</tbody>
</table>

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 → 45.43) when $\epsilon$ 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 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 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 → 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\textsubscript{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.

---

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 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.

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


[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


[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


[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


[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


[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


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


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

[69] Siamak Shakier, Cicero Nogueira dos Santos, Henry Zhu, Patrick Ng, Feng Nan, Zhiqiao 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


[77] 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


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