OpenViDial: A Large-Scale, Open-Domain Dialogue Dataset with Visual Contexts

Yuxian Meng[♣], Shuhe Wang[♠], Qinghong Han[♣]
Xiaofei Sun[♣], Fei Wu[♦], Rui Yan[★] and Jiwei Li[♦♣]
[♦]Zhejiang University, [♠]Computer Center of Peking University
[★]Gaoling School of Artificial Intelligence, Renmin University of China
[♣] Shannon.AI
{yuxian_meng, qinghong_han, xiaofei_sun, jiwei_li}@shannonai.com
wangshuhe@stu.pku.edu.cn, wufei@zju.edu.cn, ruiyan@ruc.edu.cn

Abstract

When humans converse, what a speaker will say next significantly depends on what he sees. Unfortunately, existing dialogue models generate dialogue utterances only based on preceding textual contexts, and visual contexts are rarely considered. This is due to a lack of a large-scale multi-module dialogue dataset with utterances paired with visual contexts.

In this paper, we release **OpenViDial**, a large-scale multi-module dialogue dataset. The dialogue turns and visual contexts are extracted from movies and TV series, where each dialogue turn is paired with the corresponding visual context in which it takes place. OpenViDial contains a total number of 1.1 million dialogue turns, and thus 1.1 million visual contexts stored in images.

Based on this dataset, we propose a family of encoder-decoder models leveraging both textual and visual contexts, from coarse-grained image features extracted from CNNs to fine-grained object features extracted from Faster R-CNNs. We observe that visual information significantly improves dialogue generation qualities, verifying the necessity of integrating multi-modal features for dialogue learning. Our work marks an important step towards large-scale multi-modal dialogue learning.[1]

1 Introduction

Giving machines the ability to converse like humans in the open domain is a key point towards passing the Turing test (Turing, 2009), and developing open-domain dialogue agents is of growing interest (Li et al., 2017; Ghazvininejad et al., 2017; Zhou et al., 2017; Gao et al., 2018; Asghar et al., 2018; Zhou et al., 2020). Existing approaches towards developing open-domain dialogue agents are mostly data-driven, for which a large-scale dataset is first collected. The dataset usually consists of millions of turns of dialogue utterances from real human conversations. A neural model is then trained on the dataset, learning to predict the upcoming dialogue turn conditioned on the previous textual contexts. (Li et al., 2016b,a; Zhang et al., 2018; Huang et al., 2020)

Figure 1: Two examples drawn from OpenViDial showing the necessity of considering visual contexts for dialogues.

One important aspect that existing open-domain dialogue models miss is the consideration of multi-modal features in dialogue, especially visual features. When humans converse, what a speaker should say next significantly depends on what he sees. The granularity of visual features could be as large as the location that a conversation takes place in (e.g., a cafeteria or a theater), or as small as his dialogue partner's facial expressions. For example, in Figure 1, we present two short conversations where visual contexts are crucial. In both examples, if the model has no access to visual information, it is hard to correctly generate dialogue utterances "*see the picture*" and "*moving to the attic*" in response to the preceding contexts. Unfortunately, existing dialogue models generate dialogue utter-

[1]Dataset, visual features and code are found at https://github.com/ShannonAI/OpenViDial.

ances only based on preceding textual contexts and no visual contexts are considered. This is because of the lack of a large-scale multi-modal dialogue dataset with utterances paired with visual context.

Visual dialogue (Das et al., 2017a; Lu et al., 2017; Seo et al., 2017; Wu et al., 2018; Schwartz et al., 2019; Jiang et al., 2019) is actually not a new concept. For the Visual Dialogue (VisDial) dataset (Das et al., 2017a), an image is presented to Turkers. A dialogue agent asks questions about the image, e.g., *How many people are in wheelchairs*, and the other dialogue agent gives answer such as *two*. This process is repeated for 10 rounds. Video dialog (Hori et al., 2018; Nguyen et al., 2018; Kumar et al., 2018; Lin et al., 2019; Alamri et al., 2019) takes a step further, where instead of an image, a short video is presented to Turkers, who are asked to do 10 rounds of question answering regarding the video. Important as existing visual dialogue datasets, they can be viewed as extension of visual question answering (VQA) (Antol et al., 2015; Zhang et al., 2016; Lu et al., 2016; Xu and Saenko, 2016; Goyal et al., 2017; Anderson et al., 2018; Cadene et al., 2019), which focuses more on question answering regarding an image or video, rather than dialogue learning. QA-focused dialogue datasets do not really mimic the situations where common dialogues happen. Additionally, since they focus more on generating the correct answers to questions, important aspects for dialogue generation such as coherence, diversity, informativeness and interestingness cannot be modeled or measured.

In this paper, we collect and release OpenViDial, a large-scale open-domain dialogue dataset with visual contexts. The dialogue turns and visual contexts are extracted from movies and TV series, where each dialogue turn is paired with the corresponding visual context in which it takes place. OpenViDial contains a total number of 1.1 million dialogue turns, and thus 1.1 million of visual contexts stored in images.

Using OpenViDial, we formalized the task of multi-modal dialogue generation as predicting the upcoming turn of dialogue utterances given preceding textual and visual contexts. Based on the dataset, we propose a family of encoder-decoder models leveraging both textual and visual contexts to predict the next dialogue turn. We explore different ways to incorporate visual context features into the encoder-decoder framework, including : (1) extracting global but corase-grained image features usings CNNs; and (2) extracting local but fine-grained object features using Faster R-CNNs (Ren et al., 2015). We observe progressive performance boosts as we move from coarse-grained to fined-grained image features, verifying the necessity of integrating multi-modal features for dialogue learning. Our work marks an important step towards large-scale multi-modal dialogue learning, and we wish this work will encourage more researches into building multi-modal models for open-domain dialogue generation.

The contributions of this work can be summarized as follows:

- We release OpenViDial, a large-scale multi-modal open-domain dialogue dataset, where each dialogue utterance is paired with an image of the visual context.
- We propose a line of models to leverage both long-term contextual features and visual features for multi-modal dialogue generation by incorporating both coarse-grained global image features and fine-grained local visual object features, based on which we observe consistent performance boosts. This verifies the significant merits of integrating multi-modal features for dialogue learning.

The rest of this paper is organized as follows: related work is described in Section 2. The details of dataset construction are shown in Section 3. The proposed models are prensent in Section 4. Experimental results are detailed in Section 5, followed by a brief conclusion in Section 6.

2 Related Work

2.1 Existing Dialog Datasets

Open Domain Dialog Datasets Over the past few years, various open-domain dialog datasets have been developed. The OpenSubtitle dataset (Tiedemann, 2009, 2012; Lison and Tiedemann, 2016) consists of large-scale movie conversations extracted from the OpenSubtitle website. It includes a total number of 1,782 bitexts with 3.35G sentence fragments. The Twitter Triple Corpus (Sordoni et al., 2015) consists of 4,232 Twitter conversation triples evaluated from 33K candidate triples by human raters, with 2,118 triples as tuning set and 2,114 as test set. The Cornell Movie-Dialogs Corpus (Danescu-Niculescu-

Mizil and Lee, 2011) contains a collection of fictional conversations extracted from raw movie scripts. Other plain-text dialog datasets include the Ubuntu Dialog Corpus (Lowe et al., 2015), PersonaChat (Zhang et al., 2018), EmpatheticDialogues (Rashkin et al., 2018), etc. The datasets described above only consist of texts in the form of dialogues, with no visual information included.

Visual Dialog Datasets The task of Visual Dialog is first introduced by Das et al. (2017a), where a model is required to answer a series of questions grounded in an image, given a dialog history and the image itself as contexts. Further, Das et al. (2017a) released the VisDial v0.9 and v1.0 datasets as benchmarks. The v1.0 dataset contains 120K images from MS COCO[2] and each image is associated with 10 rounds of question-answer dialog, making up 1.2M examples in total. The GuessWhat?! dataset (de Vries et al., 2017) focuses on high-level image understanding and is more goal-oriented: models need to locate an unknown object in an informative image scene by answering a sequence of "yes or no" questions. The CLEVER-Dialog (Kottur et al., 2019) and MNIST-Dialog (Seo et al., 2017) datasets are developed for diagnostic purposes. They are crafted to test the reasoning capability of visual dialog models based on the image and prior dialog turns. More recently, the Audio Visual Scene-Aware Dialog (AVSD) dataset (Hori et al., 2018; Alamri et al., 2019) was introduced. It contains more than 11,000 conversations paired with videos of human-centered activities, serving as a benchmark for the scene-aware video dialog task. The datasets described above mainly focus on answering questions regarding an image or video, and thus are more concerned about question answering rather than dialogue generation.

2.2 Dialogue Generation Models

Open Domain Dialog Generation Building open-domain dialog systems that can converse with humans has a long history in natural language processing (Weizenbaum, 1966; COLBY, 1975; Wallace, 2009). Recent advances of neural networks have spurred great interests in developing neural-based data-driven dialog models (Vinyals and Le, 2015; Li et al., 2015; Dodge et al., 2016; Serban et al., 2016; Zhao et al., 2017; Xie et al., 2017; Lee et al., 2019; Ghandeharioun et al., 2019; Li, 2020; Han et al., 2020; Zhang et al., 2019; Roller et al., 2020). Built on top of sequence-to-sequence frameworks (Sutskever et al., 2014; Vaswani et al., 2017b), neural-based dialog models are able to generate coherent (Li et al., 2016b, 2017; Tian et al., 2017; Bosselut et al., 2018; Adiwardana et al., 2020), diverse (Xu et al., 2018; Baheti et al., 2018; Tao et al., 2018), personalized (Li et al., 2016a; Luan et al., 2017; Zhang et al., 2018; Zheng et al., 2019a,b; Madotto et al., 2019), informative (Shao et al., 2017; Lewis et al., 2017; Ghazvininejad et al., 2017; Young et al., 2017; Zhao et al., 2019) and knowledge-fused (Hua et al., 2020; Zhao et al., 2020; He et al., 2020) responses, as well as bias toward different specific attributes or topics (Xing et al., 2016; Zhou et al., 2017; Wang et al., 2017; Niu and Bansal, 2018; See et al., 2019).

Visual Dialog Generation Since natural utterances and visual images are in different modalities, attention mechanisms to model the interplay between conversational utterances and visual contents are widely used (Lu et al., 2017; Kottur et al., 2018; Jiang et al., 2019; Yang et al., 2019; Guo et al., 2019; Niu et al., 2019; Kang et al., 2019; Park et al., 2020; Jiang et al., 2020b). Seo et al. (2017) employed memories to store (attention, key) pairs that can be used to retrieve the most relevant attention maps for the current question in text. Schwartz et al. (2019) designed the factor graph attention model to connect an arbitrary number of modalities with attention flows. Gan et al. (2019) proposed ReDAN, a recurrent dual attention network enhanced by a multi-step reasoning mechanism. Techniques such as reinforcement learning (Das et al., 2017b; Wu et al., 2018), variational auto-encoders (Massiceti et al., 2018) and graph networks (Zheng et al., 2019c; Jiang et al., 2020a) have also been applied to deal with the visual dialog task. Empowered by large-scale pretraining techniques, pretraining based models have made promising progress (Lu et al., 2019; Tan and Bansal, 2019; Su et al., 2019; Alberti et al., 2019; Li et al., 2019a,b; Chen et al., 2019; Wang et al., 2020; Li et al., 2020), significantly boosting the performances in terms of different metrics.

3 Constructing OpenViDial

In this section, we describe the details for OpenViDial construction. The main idea of dataset generation is to pair conversation scripts with images in movies or TV series, and use these images as visual contexts for dialogue learning.

[2] http://mscoco.org/

Number of turns	1.1M
Number of images	1.1M
Vocab size before BPE	70K
Vocab size after BPE	30K
Average length of each episode	14
Average length of each turn	7.6

Table 1: Detailed statistics for OpenViDial

We collect a raw dataset containing English movies and TV series with a total length of roughly 8,000 hours. Each second of videos can be further divided into 20∼40 frames, where each frame is an image.

3.1 Subtitle Extraction based on OCR

Because of the fact that only a small proportion of movies readily come with subtitle files, and that for most movies, subtitles are embedded in images, we need to build models to extract conversation scripts from images. To build a conversation dataset with millions of turns of image-text pairs, it is prohibitively expensive and time-intensive to employ human labors to separate each image frame with embedded scripts. We thus rely on the technique of optical character recognition (OCR) for automatic extraction of conversation subtitles from movie images.[3] We tailor the OCR model to the task of subtitle extraction, and achieves an almost perfect accuracy.

Existing open-sourced OCR models are not fit for our purpose since they are not tailored to subtitle extraction in the context of movies and TV series. We thus need to train our own OCR model.

Training Data Generation We first synthesize the OCR training dataset, where we embed texts into images to form training examples. To achieve this goal, we first need to collect text-free images from raw videos, to which texts will be later added. This is done by running an existing open-sourced OCR model[4] on video images, and pick images with no text character identified by the model. Since at this stage, our goal of identifying whether an image contains text character is a relatively easy task[5], a super accurate OCR model is not necessary and the open-sourced OCR model suffices to fulfill our need. With text-free images in hand, we pair them with texts. Texts are randomly selected from the CommonCrawl English corpus, then added to the images. Texts in images are generated using different fonts[6] and sizes. We generated a dataset containing about 10M images paired with texts.

Model Training Standard OCR training involves two stages, the *detection* of the bounding box of texts, and the *recognition* of characters. For detection, we use the PSE model as the backbone (Wang et al., 2019), which is built upon the FPN model (He et al., 2016b) with ResNet pre-trained on ImageNet dataset. For recognition, we use the Convolutional Recurrent Neural Network (CRNN) model (Shi et al., 2016) as the backbone. We omit the details since the discussion on training OCR models is beyond the scope of this paper. We use a held-out dataset for evaluation, and the trained OCR model gets an accuracy higher than 99.98% at character level and 98.4% at the image/sentence level.

Post Processing The trained OCR model is applied to videos and TV series for script extraction. Since each second of the video consists of 20∼40 frames, most of which are nearly identical, we pick 3 frames for each second and discard the rest. We also construct an English vocabulary with top 200,000 words by frequency using a part of the CommonCrawl dataset, and remove images with unknown word from the vocabulary. This further helps us remove the influence from incorrect characters by the OCR model. In addition, the following scenarios need to be handled: (1) there are cases where a consecutive number of images are paired with the same texts. We only preserve the middle image and abandon the rest; (2) There are cases where a full dialogue turn is truncated into multiple consecutive images, with each image containing only part of the text in that dialogue turn. We train a simple discriminative model to identify whether a word in a context is the end of a sentence. Using this model, we merge texts from multiple images into a single turn and pair the text with the middle image.

3.2 Statistics for OpenViDial

We collect a final dataset of 1.1M turns, where each turn consists of a sequence of words and an

[3] An alternative is to extract scripts from audios. We find extracting scripts using OCR from images obtains a much higher accuracy than speech recognition from audios. We thus adopt the former strategy.

[4] https://github.com/JaidedAI/EasyOCR

[5] This task can be make even easier by sacrificing recall (images without characters) for precision, by making sure that all selected images do not contain characters.

[6] https://www.myfonts.com/WhatTheFont/

Dataset	Genre	Multi-Modal?	# Sentences	# Images
OpenSubtitles 2016 (Lison and Tiedemann, 2016)	Plain-text Dialog	✗	337M	–
Cornell Movie-Dialogs (Danescu-Niculescu-Mizil and Lee, 2011)	Plain-text Dialog	✗	0.3M	–
VisDial v1.0 (Das et al., 2017a)	VQA	✓	2.4M	120K
Guess-What?! (de Vries et al., 2017)	VQA	✓	0.8M	66K
AVSD (Alamri et al., 2019)	VQA	✓	152K	–
OpenViDial (this work)	Visual+Text Dialog	✓	1.1M	1.1M

Table 2: A comparison of different datasets. VQA: Visual Question Answering.

image. The size of the image is either 1280×720 or 1920×1080 based on different video resources. We employ the BPE tokenizer (Sennrich et al., 2016) for text processing. The detailed statistics for OpenViDial are shown in Table 1. We split the dataset into 1M/50K/50K for training, dev and test.

Table 2 shows the comparison between different datasets. Comparing against OpenSubtitles (Lison and Tiedemann, 2016), OpenViDial has fewer sentences but contains multi-modal features. Additionally, the OpenSubtitles dataset is an extremely noisy dataset, where consecutive lines may not appear in the same conversation or scene, and may not even be spoken by the same character. Comparing with other datasets with visual features, i.e., VisDial, Guess-What?! and AVSD, OpenViDial focuses more on dialogue learning rather than question answering.

4 Models

4.1 Notations

OpenViDial consists of a set of dialogue episodes $(X, Z) \in \mathcal{D}$. $X = \{x_1, \cdots, x_n, \cdots\}$ is a sequence of dialog turns in texts and $Z = \{z_1, \cdots, z_n, \cdots\}$ is a sequence of images. z_n and x_n are paired, in which z_n denotes the visual context which x_n takes place in. n denotes the length of the dialog episode. Each dialog turn $x_j (1 \leq j \leq n)$ is a sequence of tokens, where $x_j = \{w_{j,1}, w_{j,2}, ..., w_{j,n_j}\}$, and n_j denotes the length of the text turn x_j. $h_{j,k}$ denotes the word representation for token $w_{j,k}$. A model is required to generate the next dialog utterance x_{j+1} conditioning on all preceding (x, z) pairs along with the visual information z_{j+1} for the current dialog turn, i.e., maximizing $p(x_{j+1}|x_{\leq j}, z_{\leq j+1})$. Below we present three models to handle this conditional probability.

4.2 Model 1: NoVisual (NV)

The most naive model is to use only dialog texts without visual information, where the model degenerates to a plain-text dialog generation model. The model is optimized to minimize the following negative log-likelihood (NLL) loss:

$$\mathcal{L}_{\text{NoVisual}} = -\sum_{(X,Z)\in\mathcal{D}} \sum_{j=0}^{n-1} p(x_{j+1}|x_{\leq j}) \quad (1)$$

We use a standard Transformer architecture (Vaswani et al., 2017a) as the model backbone. The numbers of encoder layer and decoder layer are both 3, with 5 heads in each layer and input dimension of 512. We pack the preceding (up to 5) dialog history $x_{\leq j}$ into a long sequence with a spacial `[SEP]` token as the delimiter between two consecutive dialog turns. An absolute positional embedding is added to each word representation (Devlin et al., 2018) .

4.3 Model 2: CoarseVisual (CV)

Our second model employs a naive approach to inject visual information into dialog generation, which we refer to as CoarseVisual (CV). More concretely, we first use a pre-trained ResNet-50 (He et al., 2016a) on ImageNet to induce a high-dimensional feature f_j for image z_j. Then, for all tokens $w_{j,k}$ in the j-th dialog utterance, we add the image feature f_j to its word representation $h_{j,k}$, forming the input layer representation $h^0_{j,k}$ as the input to the encoder-decoder model:

$$h^0_{j,k} = h_{j,k} + f_j \quad (2)$$

The concatenation of all input token representations for the j-th dialog utterance is denoted by $h^0_j = [h^0_{j,1}, \cdots, h^0_{j,n_j}]$. Hence, the input to the encoder is given by $\{\texttt{[CLS]}, h^0_1, \texttt{[SEP]}, h^0_2, \texttt{[SEP]}, \cdots, h^0_j, \texttt{[SEP]}, f_{j+1}, \texttt{[SEP]}\}$. f_{j+1} represents the encoded feature of image z_{j+1}. Afterwards, the CoarseVisual model is learning to generate the forthcoming dialog utterance x_{j+1} by minimizing the following

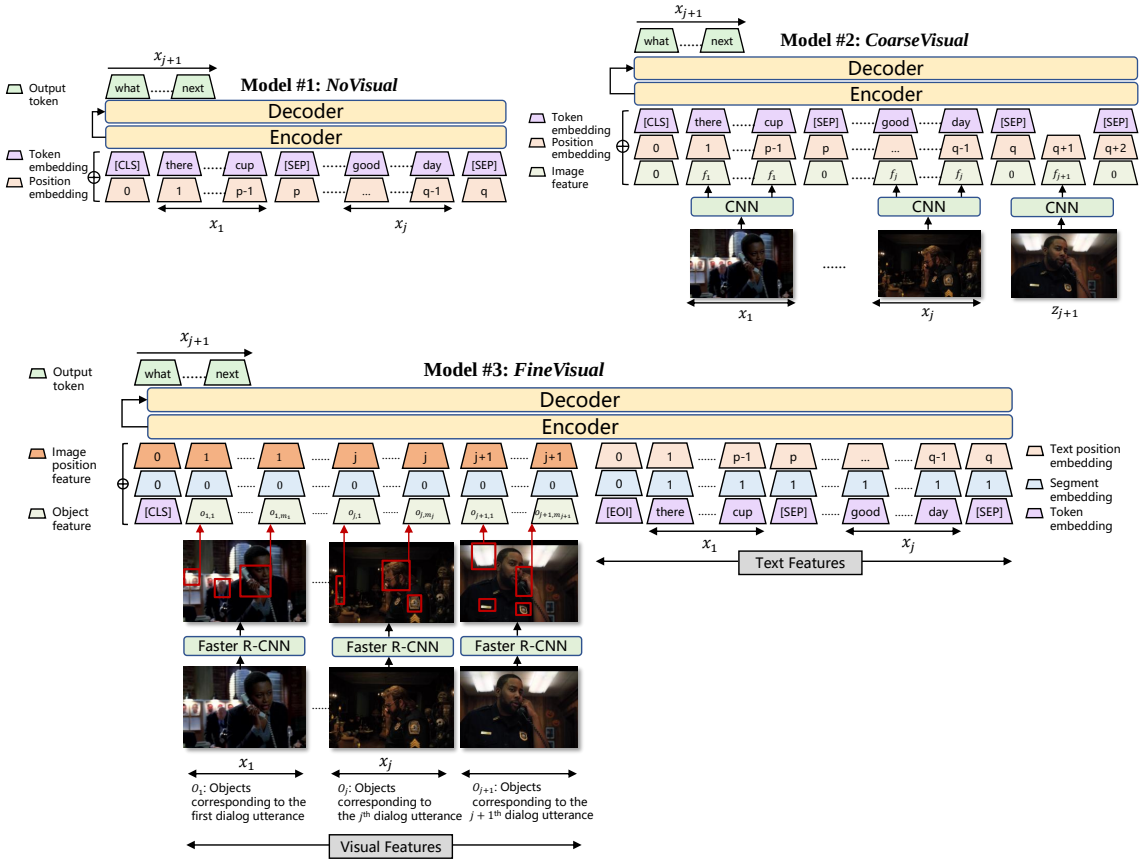


Figure 2: An overview of the proposed models NoVisual, CoarseVisual and FineVisual.

NLL loss:

$$\mathcal{L}_{\text{CoarseVisual}} = -\sum_{(X,Z)\in\mathcal{D}}\sum_{j=0}^{n-1} p(x_{j+1}|x_{\leq j}, f_{\leq j+1}) \tag{3}$$

Again, we use $\text{Transformer}_{\text{base}}$ as the backbone.

4.4 Model 3: FineVisual (FV)

While the CoarseVisual model is able to combine the vision and text modalities, it performs at a coarse level for extracting global image features. This might be insufficient to model fine-grained visual elements in images such as facial expressions, body gestures as well as physical motions. Hence, we use Faster R-CNN (Ren et al., 2015) pretrained on Visual Genome (Krishna et al., 2017) to extract fine-grained visual semantic objects. For an input image z_j, Faster R-CNN returns a set of detected objects in the image, each of which is captured by a dense feature representation. Let $O_j = \{o_{j,1}, \cdots, o_{j,q}, \cdots, o_{j,m_j}\}$ denote the set of extracted features for image z_j, where m_j is the number of extracted objects for z_j. Each extracted feature can be mapped back to a bounding box / region (i.e., Region-of-Interest (RoI)) in the original image. For each dialog turn x_{j+1} to generate, the input to the model is $\{\texttt{[CLS]}, O_1, \cdots, O_{j+1}, \texttt{[EOI]}, x_1, \texttt{[SEP]}, \cdots, x_j, \texttt{[SEP]}\}$. `[EOI]` is a special end-of-image token denoting the end of the sequence of object features. Similar to the CoarseVisual model, the FineVisual model is optimized to minimize the following NLL loss:

$$\mathcal{L}_{\text{FineVisual}} = -\sum_{(X,Z)\in\mathcal{D}}\sum_{j=0}^{n-1} p(x_{j+1}|O_{\leq j+1}, x_{\leq j}) \tag{4}$$

The input to the encoder consists of two parts: text features i.e. $\{x_1, \texttt{[SEP]}, \cdots, x_j, \texttt{[SEP]}\}$, and visual features $\{O_1, \cdots, O_{j+1}, \texttt{[EOI]}\}$. For visual features, image-specific position feature highlighting objects across different images is added to object representations. To further distinguish the visual part from the text part, we respectively add an image-notifying embedding to all visual objects and a text-notifying embedding to all text tokens. An illustration of the three models is shown in Figure 2.

Model	BLEU-1	BLEU-2	BLEU-4	Stopword%
NV	14.01	3.98	1.07	58.1%
CV	14.58	4.35	1.14	54.2%
FV	**15.61**	**4.71**	**1.22**	**52.9%**

Table 3: Automatic evaluation results for BLEU and Stopword%.

Model	Dis-1	Dis-2	Dis-3	Dis-4
NV	0.0091	0.0355	0.0682	0.1018
CV	0.0108 (+18.7%)	0.0448 (+26.2%)	0.0915 (+34.2%)	0.1465 (+43.9%)
FV	**0.0118 (+29.7%)**	**0.0502 (+41.4%)**	**0.1082 (+58.7%)**	**0.1778 (+74.7%)**

Table 4: Automatic evaluation results for diversity.

4.5 Training

For all models, we use Adam (Kingma and Ba, 2014) with learning rate of 1e-4, β_1 = 0.9, β_2 = 0.999, warmup over the first 10K steps, and linear decay of the learning rate. We use a dropout rate of 0.1 on all layers (including the softmax layer).

5 Experiments

5.1 Automatic Evaluation

We use the following metrics for automatic evaluation:

- **BLEU**: Following Li et al. (2015); Sordoni et al. (2015), we report BLEU scores for evaluation, which measure the n-gram overlaps between the generated sequences and golden target sequences. We use $n = 1, 2, 4$.
- **Diversity**: Following Li et al. (2015), we report the degree of diversity by calculating the number of distinct n-grams (Dis-n, $n = 1, 2, 3, 4$) in generated responses. The value is scaled by the total number of generated tokens to avoid favoring long sentences.
- **Stopword%**: Stopword% is the percentage of stop-words[7] and punctuations of the responses generated by each model.

Results are shown in Table 3 and Table 4. As we can see, for both BLEU and diversity, we observe progressive performance boosts from NoVisual to CoarseVisual, then to FineVisual. the performance of the FineVisual model is significantly better than that of the NoVisual and CoarseVisual models, indicating that integrating more fine-grained visual features leads to better multi-modal dialogue learning abilities. The percentages of stopwords for NV, CV and FC are respectively 58.1%, 51.5% and 52.9%. These results show the effectiveness of generating more informative responses by injecting visual contexts.

[7] https://www.ranks.nl/stopwords

5.2 Adversarial Evaluation

The adversarial evaluation strategy is proposed by Kannan and Vinyals (2017); Li et al. (2017) to train a discriminator function to label dialogues as machine-generated (negative) or human-generated (positive). Positive examples are taken from training dialogues, while negative examples are decoded from a model.

The input to the discriminator is the concatenation of features for constituent dialogue turns, including the preceding features and the generated text. For each dialogue turn, the feature includes both visual features extracted from the image using CNNs, and text features using word embeddings. A multi-layer transformer is built on top of the concatenation, with the `[CLS]` feature fed to the sigmoid function, the output of which denotes the probability of whether the generated text is machine-generated or human-generated. We used examples from the dev set to train the discriminator, in which we treat half of the examples with original responses in the dataset as human-generated, and the other half with model generated responses as machine-generated. We test the trained model on the test set generated in the same way. We report adversarial success, which is the percentage of the generated responses that can fool the evaluator to believe that it is human-generated, equal to $1 - accuracy$, where $accuracy$ denotes the accuracy of the performance of the discriminator on the test set. Higher values of adversarial success indicates better dialogue generation models.

NV, CV and FV respectively obtain adversarial success values of 0.942, 0.917 and 0.884, demonstrating that integrating visual contexts facilitate generating responses more mimicking human conversations.

5.3 Human Evaluation

Both automatic evaluations and adversarial evaluation suffer from disadvantages. For the former, there have been debates on their validity for dia-

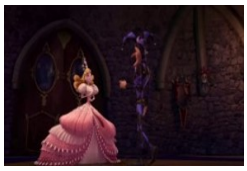
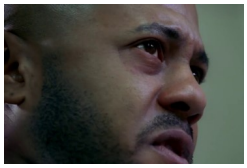
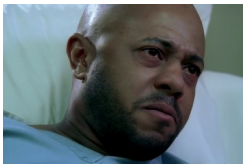
Time Step $t-2$	**Time Step** $t-1$	**Prediction (Time Step** t**)**
		
Context: No .	*Context*: The Land of OZ is already falling into ruin because of you .	*NV*: I'm sorry . I'm sorry . *CV*: I'm sorry . I'm sorry . I'm sorry . *FV*: You are the queen of the castle . *Truth*: No good can come from the reign of a fool .
		
Context: I wouldn't be asking you if it wasn ' t an emergency .	*Context*: Please , just let me use that phone .	*NV*: I don ' t know what you ' re talking about . *CV*: I ' m sorry , sir . *FV*: I need you to call me back . *Truth*: Agent Lang ?

Table 5: Randomly chosen examples from the test set.

logue generation (Lowe et al., 2018); for the latter, it requires training another model (i.e., the discriminator) for evaluation. We thus conduct human evaluation for further validations. We employ crowdsourced judges to provide evaluations for a random sample of 1000 episodes from the test set. For each input context, we present annotators with both preceding text contexts, preceding visual contexts, and the current visual context, along with outputs from the three models, i.e., NV, CV and FV. Annotators were asked to score every model response on a 5-point scale (Strongly Agree, Agree, Unsure, Disagree, Strongly Disagree) based on three aspects: *Relevance* (whether the generated response is relevant to the contexts, both visual and textual), *Diversity* (whether the generated response has diverse words) and *Readability* (whether the generated response is grammatical). Ratings were later collapsed to 3 categories (Agree, Unsure, Disagree). Additionally, to rule out the impact from visual contexts, we conduct another setup where visual contexts are hidden from the annotators, with only texts presented. Results are shown in Table 6.

To verify the statistical significance of the reported results, we perform a pairwise bootstrap test (Johnson, 2001; Berg-Kirkpatrick et al., 2012) to compare the difference between the percentage of responses that are labeled as "yes". We find that

Model	**No** %	**Unsure** %	**Yes** %
With Visual			
NV	34.9	27.2	37.9
CV	25.5	25.6	49.8
FV	21.5	25.1	53.4
Without Visual			
NV	31.2	28.5	40.3
CV	28.6	26.9	44.5
FV	23.8	25.5	50.7

Table 6: Human evaluation results.

FV is significantly better than CV, which is significantly better than NV, with p-value < 0.01. This validates the importance of harnessing visual contexts for dialogue generation. Interestingly, even hiding visual information, annotations still think outputs from FV and CV are of better quality than NV. This is because visual contexts provide grounding information for dialogue to proceed and avoid dull responses.

5.4 Examples

In Table 5, we present two randomly chosen examples from the test set. Contents in "Time Step $t-1$" and "Time Step $t-2$" correspond to contexts of text utterances and images in preceding time steps. Contents in "Prediction (Time Step t)"

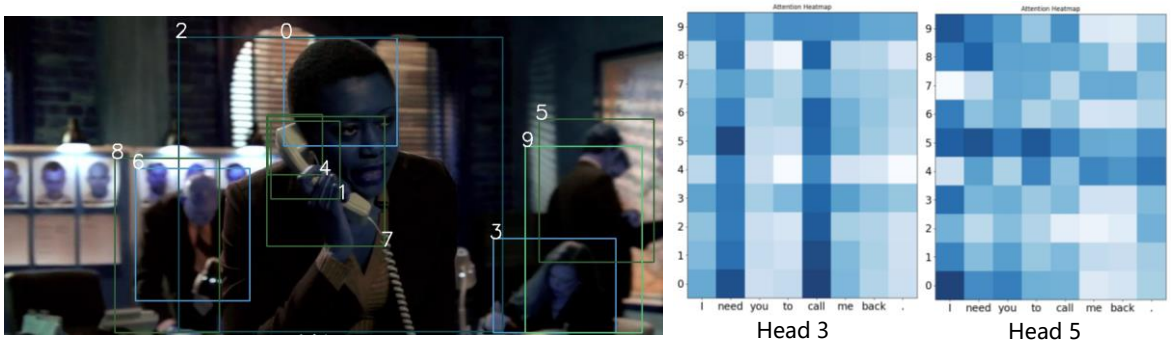


Figure 3: Attention weight visualization in the FV model for a sampled example.

shows generated responses from different models as well as the ground-truth responses. We can see that the NoVisual model tends to generate dull and meaningless responses such as "I'm sorry" or "I don't know", since it has no access to the visual information and thus does not know how to make the conversation proceed. This inevitably results in irrelevant and dull responses. Instead, the other two models, particularly the FineVisual model, can generate more informative and relevant responses due to their ability to digest visual contexts. For instance, in the first example, FV recognizes a cartoon queen in the image and produces "You are the queen of the castle" as a response.

In Figure 3, we visualize the attention weights for generating the response "*I need you to call me back*" w.r.t. the top 10 detected objects from Faster R-CNNs. We can observe that in the attention heatmap, the model is able to capture the correlation between the generated word "call" and the man who is calling (object 0, 1, 2). In the attention heatmap, the model attaches more attentions to the person in the image (object 0, 3, 5, 6, 9) when generating the word "I". These results show that the proposed FV model is able to model multi-modal interactions between texts and images through self-attentions and thus generate meaningful responses relevant to the visual contexts.

6 Conclusion

In this paper, we release OpenViDial, a large-scale open-domain dialogue dataset with visual contexts. In OpenViDial, each dialogue turn is paired with the corresponding visual context in which it takes place. Based on OpenViDial, we propose a family of encoder-decoder models leveraging both textual and visual contexts for better dialogue generation. Our work marks an important step towards large-scale multi-modal dialogue learning.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. 2019. Audio visual scene-aware dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7558–7567.

Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. 2019. Fusion of detected objects in text for visual question answering. *arXiv preprint arXiv:1908.05054*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Nabiha Asghar, Pascal Poupart, Jesse Hoey, Xin Jiang, and Lili Mou. 2018. Affective neural response generation. In *European Conference on Information Retrieval*, pages 154–166. Springer.

Ashutosh Baheti, Alan Ritter, Jiwei Li, and Bill Dolan. 2018. Generating more interesting responses in neural conversation models with distributional constraints. *arXiv preprint arXiv:1809.01215*.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural*

Language Processing and Computational Natural Language Learning, pages 995–1005, Jeju Island, Korea.

Antoine Bosselut, Asli Celikyilmaz, Xiaodong He, Jianfeng Gao, Po-Sen Huang, and Yejin Choi. 2018. Discourse-aware neural rewards for coherent text generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 173–184, New Orleans, Louisiana. Association for Computational Linguistics.

Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. 2019. Murel: Multimodal relational reasoning for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1989–1998.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*.

KENNETH MARK COLBY. 1975. Chapter 4 - language-recognition processes for understanding dialogues in teletyped psychiatric interviews. In KENNETH MARK COLBY, editor, *Artificial Paranoia*, pages 37 – 49. Pergamon.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017a. Visual dialog.

Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. 2017b. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of the IEEE international conference on computer vision*, pages 2951–2960.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jesse Dodge, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander Miller, Arthur Szlam, and Jason Weston. 2016. Evaluating prerequisite qualities for learning end-to-end dialog systems.

Zhe Gan, Yu Cheng, Ahmed El Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. 2019. Multi-step reasoning via recurrent dual attention for visual dialog. *arXiv preprint arXiv:1902.00579*.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational ai. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1371–1374.

Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard. 2019. Approximating interactive human evaluation with self-play for open-domain dialog systems. In *Advances in Neural Information Processing Systems*, pages 13658–13669.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2017. A knowledge-grounded neural conversation model. *arXiv preprint arXiv:1702.01932*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Dan Guo, Hui Wang, and Meng Wang. 2019. Dual visual attention network for visual dialog. In *IJCAI*, pages 4989–4995.

Qinghong Han, Yuxian Meng, Fei Wu, and Jiwei Li. 2020. Non-autoregressive neural dialogue generation.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016a. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016b. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer.

Wanwei He, Min Yang, Rui Yan, Chengming Li, Ying Shen, and Ruifeng Xu. 2020. Amalgamating knowledge from two teachers for task-oriented dialogue system with adversarial training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3498–3507, Online. Association for Computational Linguistics.

Chiori Hori, Huda Alamri, Jue Wang, Gordon Wichern, Takaaki Hori, Anoop Cherian, Tim K. Marks, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, Irfan Essa, Dhruv Batra, and Devi Parikh. 2018. End-to-end audio visual scene-aware dialog using multimodal attention-based video features.

Kai Hua, Zhiyuan Feng, Chongyang Tao, Rui Yan, and Lu Zhang. 2020. Learning to detect relevant contexts and knowledge for response selection in retrieval-based dialogue systems. In *Proceedings of the 29th ACM International Conference on Information Knowledge Management*, CIKM '20, page 525–534, New York, NY, USA. Association for Computing Machinery.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32.

Xiaoze Jiang, Siyi Du, Zengchang Qin, Yajing Sun, and Jing Yu. 2020a. 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.

Xiaoze Jiang, Jing Yu, Zengchang Qin, Yingying Zhuang, Xingxing Zhang, Yue Hu, and Qi Wu. 2019. Dualvd: An adaptive dual encoding model for deep visual understanding in visual dialogue.

Xiaoze Jiang, Jing Yu, Yajing Sun, Zengchang Qin, Zihao Zhu, Yue Hu, and Qi Wu. 2020b. Dam: Deliberation, abandon and memory networks for generating detailed and non-repetitive responses in visual dialogue.

Roger W. Johnson. 2001. An introduction to the bootstrap. *Teaching Statistics*, 23(2):49–54.

Gi-Cheon Kang, Jaeseo Lim, and Byoung-Tak Zhang. 2019. Dual attention networks for visual reference resolution in visual dialog. *arXiv preprint arXiv:1902.09368*.

Anjuli Kannan and Oriol Vinyals. 2017. Adversarial evaluation of dialogue models. *arXiv preprint arXiv:1701.08198*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2018. Visual coreference resolution in visual dialog using neural module networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 153–169.

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

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Shachi H Kumar, Eda Okur, Saurav Sahay, Juan Jose Alvarado Leanos, Jonathan Huang, and Lama Nachman. 2018. Context, attention and audio feature explorations for audio visual scene-aware dialog. *arXiv preprint arXiv:1812.08407*.

Sungjin Lee, Qi Zhu, Ryuichi Takanobu, Xiang Li, Yaoqin Zhang, Zheng Zhang, Jinchao Li, Baolin Peng, Xiujun Li, Minlie Huang, et al. 2019. Convlab: Multi-domain end-to-end dialog system platform. *arXiv preprint arXiv:1904.08637*.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453, Copenhagen, Denmark. Association for Computational Linguistics.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. 2019a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training.

Jiwei Li. 2020. Teaching machines to converse. *arXiv preprint arXiv:2001.11701*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016a. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016b. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019b. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Kuan-Yen Lin, Chao-Chun Hsu, Yun-Nung Chen, and Lun-Wei Ku. 2019. Entropy-enhanced multimodal attention model for scene-aware dialogue generation. *arXiv preprint arXiv:1908.08191*.

P. Lison and J. Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *LREC*.

Ryan Lowe, Michael Noseworthy, Iulian V. Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2018. Towards an automatic turing test: Learning to evaluate dialogue responses.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

Jiasen Lu, Anitha Kannan, Jianwei Yang, Devi Parikh, and Dhruv Batra. 2017. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for vi-

sual question answering. In *Advances in neural information processing systems*, pages 289–297.

Yi Luan, Chris Brockett, Bill Dolan, Jianfeng Gao, and Michel Galley. 2017. Multi-task learning for speaker-role adaptation in neural conversation models. *arXiv preprint arXiv:1710.07388*.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5459.

Daniela Massiceti, N Siddharth, Puneet K Dokania, and Philip HS Torr. 2018. Flipdial: A generative model for two-way visual dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6097–6105.

Dat Tien Nguyen, Shikhar Sharma, Hannes Schulz, and Layla El Asri. 2018. From film to video: Multi-turn question answering with multi-modal context.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data.

Yulei Niu, Hanwang Zhang, Manli Zhang, Jianhong Zhang, Zhiwu Lu, and Ji-Rong Wen. 2019. Recursive visual attention in visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6679–6688.

Sungjin Park, Taesun Whang, Yeochan Yoon, and Hueiseok Lim. 2020. Multi-view attention networks for visual dialog. *arXiv preprint arXiv:2004.14025*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Idan Schwartz, Seunghak Yu, Tamir Hazan, and Alexander G Schwing. 2019. Factor graph attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2039–2048.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. *arXiv preprint arXiv:1902.08654*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Paul Hongsuck Seo, Andreas Lehrmann, Bohyung Han, and Leonid Sigal. 2017. Visual reference resolution using attention memory for visual dialog. In *Advances in neural information processing systems*, pages 3719–3729.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2016. A hierarchical latent variable encoder-decoder model for generating dialogues. *arXiv preprint arXiv:1605.06069*.

Louis Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. *arXiv preprint arXiv:1701.03185*.

Baoguang Shi, Xiang Bai, and Cong Yao. 2016. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Chongyang Tao, Shen Gao, Mingyue Shang, Wei Wu, Dongyan Zhao, and Rui Yan. 2018. Get the point of my utterance! learning towards effective responses with multi-head attention mechanism. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4418–4424. International Joint Conferences on Artificial Intelligence Organization.

Zhiliang Tian, Rui Yan, Lili Mou, Yiping Song, Yansong Feng, and Dongyan Zhao. 2017. How to make context more useful? an empirical study on context-aware neural conversational models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–236, Vancouver, Canada. Association for Computational Linguistics.

J. Tiedemann. 2009. News from opus — a collection of multilingual parallel corpora with tools and interfaces.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Alan M Turing. 2009. Computing machinery and intelligence. In *Parsing the turing test*, pages 23–65. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multimodal dialogue.

Richard S. Wallace. 2009. *The Anatomy of A.L.I.C.E.*, pages 181–210. Springer Netherlands, Dordrecht.

Di Wang, Nebojsa Jojic, Chris Brockett, and Eric Nyberg. 2017. Steering output style and topic in neural response generation. *arXiv preprint arXiv:1709.03010*.

Wenhai Wang, Enze Xie, Xiang Li, Wenbo Hou, Tong Lu, Gang Yu, and Shuai Shao. 2019. Shape robust text detection with progressive scale expansion network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9336–9345.

Yue Wang, Shafiq Joty, Michael R Lyu, Irwin King, Caiming Xiong, and Steven CH Hoi. 2020. Vd-bert: A unified vision and dialog transformer with bert. *arXiv preprint arXiv:2004.13278*.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Qi Wu, Peng Wang, Chunhua Shen, Ian Reid, and Anton Van Den Hengel. 2018. Are you talking to me? reasoned visual dialog generation through adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6106–6115.

Ziang Xie, Sida I Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Dan Jurafsky, and Andrew Y Ng. 2017. Data noising as smoothing in neural network language models. *arXiv preprint arXiv:1703.02573*.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2016. Topic aware neural response generation. *arXiv preprint arXiv:1606.08340*.

Huijuan Xu and Kate Saenko. 2016. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, pages 451–466. Springer.

Jingjing Xu, Xuancheng Ren, Junyang Lin, and Xu Sun. 2018. Dp-gan: diversity-promoting generative adversarial network for generating informative and diversified text. *arXiv preprint arXiv:1802.01345*.

Tianhao Yang, Zheng-Jun Zha, and Hanwang Zhang. 2019. Making history matter: History-advantage sequence training for visual dialog.

Tom Young, Erik Cambria, Iti Chaturvedi, Minlie Huang, Hao Zhou, and Subham Biswas. 2017. Augmenting end-to-end dialog systems with commonsense knowledge. *arXiv preprint arXiv:1709.05453*.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and yang: Balancing and answering binary visual questions.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Tiancheng Zhao, Kaige Xie, and Maxine Eskenazi. 2019. Rethinking action spaces for reinforcement learning in end-to-end dialog agents with latent variable models. *arXiv preprint arXiv:1902.08858*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *arXiv preprint arXiv:1703.10960*.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390, Online. Association for Computational Linguistics.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019a. Personalized dialogue generation with diversified traits. *arXiv preprint arXiv:1901.09672*.

Yinhe Zheng, Rongsheng Zhang, Xiaoxi Mao, and Minlie Huang. 2019b. A pre-training based personalized dialogue generation model with persona-sparse data.

Zilong Zheng, Wenguan Wang, Siyuan Qi, and Song-Chun Zhu. 2019c. Reasoning visual dialogs with structural and partial observations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6669–6678.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2017. Emotional chatting machine: Emotional conversation generation with internal and external memory. *arXiv preprint arXiv:1704.01074*.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.

A More Examples

In Table 7, we present more examples to show the necessity of considering visual contexts for dialog generation.

Time Step $t-2$	**Time Step** $t-1$	**Prediction (Time Step** t**)**
		
Context: Blake , I need you to go out to 59 Elm .	*Context*: There should be a guy there . New in town . Possibly shirtless .	*NV*: No , no , no , no , no . *CV*: Oh , my God . *FV*: Hey , man . I got a call for you . *Truth*:Yeah . Middle-aged , growls a lot , glue-on hairy hands ?
		
Context: Mrs. Taylor is sick .	*Context*: She has taking a casserole over to her house .	*NV*: I'm gonna to do it . *CV*: What ? *FV*: A weekly magazine . *Truth*: What are you doing ?

Table 7: Randomly chosen examples from the test set.