Multimodal Incremental Transformer with Visual Grounding for Visual Dialogue Generation

Feilong Chen, Fandong Meng, Xiuyi Chen, Peng Li, Jie Zhou
Pattern Recognition Center, WeChat AI, Tencent Inc, Beijing, China
{ivess.chan,hugheren.chan}@gmail.com
{fandongmeng,patrickpli,withtomzhou}@tencent.com

Abstract

Visual dialogue is a challenging task since it needs to answer a series of coherent questions on the basis of understanding the visual environment. Previous studies focus on the implicit exploration of multimodal co-reference by implicitly attending to spatial image features or object-level image features but neglect the importance of locating the objects explicitly in the visual content, which is associated with entities in the textual content. Therefore, in this paper we propose a **M**ultimodal **I**ncremental **T**ransformer with **V**isual **G**rounding, named MITVG, which consists of two key parts: visual grounding and multimodal incremental transformer. Visual grounding aims to explicitly locate related objects in the image guided by textual entities, which helps the model exclude the visual content that does not need attention. On the basis of visual grounding, the multimodal incremental transformer encodes the multi-turn dialogue history combined with visual scene step by step according to the order of the dialogue and then generates a contextually and visually coherent response. Experimental results on the VisDial v0.9 and v1.0 datasets demonstrate the superiority of the proposed model, which achieves comparable performance.

1 Introduction

Recently, there is increasing interest in vision-language tasks, such as image caption (Xu et al., 2015; Anderson et al., 2016, 2018; Cornia et al., 2020) and visual question answering (Ren et al., 2015a; Gao et al., 2015; Lu et al., 2016; Anderson et al., 2018). In the real world, our conversations (Chen et al., 2020b, 2019) usually have multiple turns. As an extension of conventional single-turn visual question answering, Das et al. (2017) introduce a multi-turn visual question answering task named visual dialogue, which aims to explore the ability of an AI agent to hold a meaningful multi-turn dialogue with humans in natural language about visual content.

Caption: there is a frisbee team with their coach taking a team photo

Q1: how many people ?	A1: 7 people
Q2: is anyone holding a frisbee ?	A2: yes
Q3: is the coach on the right ?	A3: yes, on the far right
Q4: are they wearing matching uniforms ?	A4: all except the coach

Figure 1: An example of visual dialogue. The color in text background corresponds to the same color box in the image, which indicates the same entity. Our model firstly associates textual entities with objects explicitly and then gives contextually and visually coherent answers to contextual questions.

Visual dialogue (Agarwal et al., 2020; Wang et al., 2020; Qi et al., 2020; Murahari et al., 2020) requires agents to give a response on the basis of understanding both visual and textual content. One of the key challenges in visual dialogue is how to solve multimodal co-reference (Das et al., 2017; Kottur et al., 2018). Therefore, some fusion-based models (Das et al., 2017) are proposed to fuse spatial image features and textual features in order to obtain a joint representation. Then attention-based models (Lu et al., 2017; Wu et al., 2018; Kottur et al., 2018) are proposed to dynamically attend to spatial image features in order to find related visual content. Furthermore, models based on object-level image features (Niu et al., 2019; Gan et al., 2019; Chen et al., 2020a; Jiang et al., 2020a; Nguyen

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 436–446

et al., 2020; Jiang et al., 2020b) are proposed to effectively leverage the visual content for multimodal co-reference. However, as implicit exploration of multimodal co-reference, these methods implicitly attend to spatial or object-level image features, which is trained with the whole model and is inevitably distracted by unnecessary visual content. Intuitively, specific mapping of objects and textual entities can reduce the noise of attention. As shown in Figure 1, the related objects can help the agent to understand the entities (e.g., Q1: "*people*", Q2: "*frisbee*", Q3: "*coach*") for the generation of correct answers. Then when it answers the question Q4 "*are they wearing matching uniforms ?*", the agent has already comprehended "*people*" and "*coach*" from the previous conversation. On this basis, it can learn the entity "*uniforms*" with the corresponding object in the image, and generate the answer "*all except the coach*". To this end, we need to 1) explicitly locate related objects guided by textual entities to exclude undesired visual content, and 2) incrementally model the multi-turn structure of the dialogue to develop a unified representation combining multi-turn utterances with the corresponding related objects. However, previous work overlooks these two important aspects.

In this paper, we thus propose a novel and effective **M**ultimodal **I**ncremental **T**ransformer with **V**isual **G**rounding, named MITVG, which contains two key parts: visual grounding and multimodal incremental transformer. Visual grounding aims to establish specific mapping of objects and textual entities by explicitly locating related objects in the image with the textual entities. By doing so, our model can exclude undesired visual content and reduce attention noise. On the basis of visual grounding, the multimodal incremental transformer is used to model the multi-turn dialogue history combined with the specific visual content to generate visually and contextually coherent responses. As an encoder-decoder framework, MITVG contains a Multimodal Incremental Transformer Encoder (MITE) and a Gated Cross-Attention Decoder (GCAD).

We test the effectiveness of our proposed model on large-scale datasets: VisDial v0.9 and v1.0 (Das et al., 2017). Both automatic and manual evaluations show that our model substantially outperforms the competitive baselines and achieves the new state-of-the-art results on substantial metrics. Our main contributions are as follows:

- To the best of our knowledge, we are the first to leverage visual grounding to explicitly locate related objects in the image guided by textual entities for visual dialogue.
- We propose a novel multimodal incremental transformer to encode the multi-turn dialogue history step by step combined with the visual content and then generate a contextually and visually coherent response.
- We achieve comparable performance on VisDial v0.9 and v1.0 datasets.

2 Approach

2.1 Overview

In this section, we formally describe the visual dialogue task and then proceed to our proposed **M**ultimodal **I**ncremental **T**ransformer with **V**isual **G**rounding (MITVG).

Following Das et al.(2017), a visual dialogue agent is given three inputs, i.e., an image I, a dialogue history (the caption and question-answer pairs) till round $t-1$: $H = (\underbrace{Cap}_{H_0}, \underbrace{(Q_1, A_1)}_{H_1}, \cdots, \underbrace{(Q_{t-1}, A_{t-1})}_{H_{t-1}})$ and the current question Q_t at round t, where Cap is the caption describing the image taken as H_0 and $H_1, \ldots, H_{t-1}$ are concatenations of question-answer pairs. The goal of the visual dialogue agent is to generate a response (or answer) A_t to the question Q_t. Cap, Q_* and A_* are sentences.

Figure 2 shows the framework of MITVG, which aims to explicitly model multi-turn dialogue history step by step based on the explicit modeling relationship between multiple modalities. MITVG firstly locates related objects in the image explicitly guided by the textual entities via visual grounding, then encodes multi-turn dialogue history in the order of the dialogue utterance based on visual grounding via Multimodal Incremental Encoder (MITE), and finally utilizes the outputs of both encoder and visual grounding to generate the response word by word via Gated Cross-Attention Decoder (GCAD).

2.2 Input Representation

Before describing our method, we introduce the input representation.

Image Features. We use a pre-trained Faster R-CNN model (Ren et al., 2015b) to extract object-

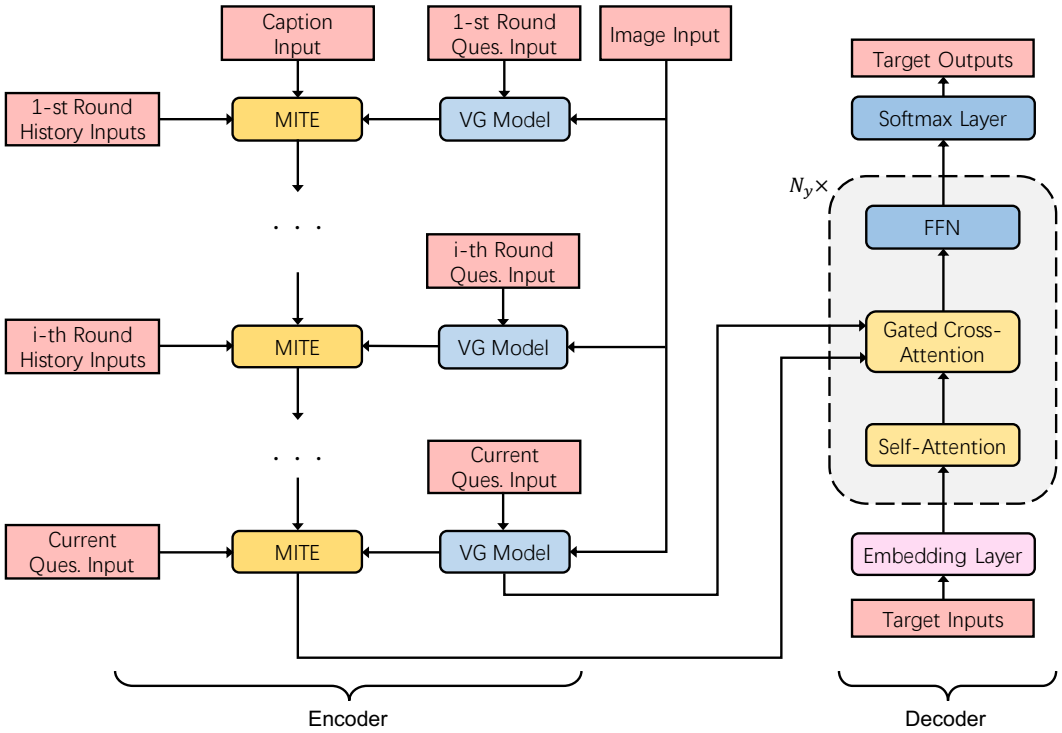

Figure 2: The framework of **M**ultimodal **I**ncremental **T**ransformer with **V**isual **G**rounding (MITVG). "VG Model" indicates visual grounding model (Yang et al., 2019b) (Details are described in Sec. 2.3). "MITE" denotes the multimodal incremental transformer encoder (Details are described in Sec. 2.4.1). MITVG firstly uses the VG model to explicitly model the relationship between the textual content and the visual content, and encodes multi-turn dialogue history in the order of the dialogue based on visual grounding, and finally utilizes the outputs of both encoder and visual grounding to generate the response word by word in the decoding process.

level image features. Specifically, the image features v for an image I are represented by:

$$v = \text{Faster R} - \text{CNN}(I) \in \mathbb{R}^{K\times V}, \quad (1)$$

where K denotes the total number of the detected objects per image and V denotes the dimension of features for each object.

Language Features. The current (at the t-th round) L-word question features are a sequence of M-dimension word embedding with positional encoding added (Vaswani et al., 2017), as follows:

$$q_t = [s_{t,1}, s_{t,2}, \ldots, s_{t,L}] \in \mathbb{R}^{L\times M}, \quad (2)$$
$$s_{t,j} = w_j + PE(j), \quad (3)$$

where w_j is the word embedding of the j-th word in the question Q_t, and $PE(\cdot)$ denotes positional encoding function (Vaswani et al., 2017). For the dialogue history $H = \{H_0, H_1, \ldots, H_{t-1}\}$ and the answer A_t, the dialogue history features $u = \{u_0, u_1, \ldots, u_{t-1}\}$ and the answer features a_t are obtained in the same way as the question Q_t.

2.3 Visual Grounding

To exclude the needless visual content, we introduce visual grounding, which is defined to ground a natural language query (phrase or sentence) about an image onto a correct region of the image. First of all, we use NeuralCoref[1] for reference resolution. For example, when it processes the question Q4 "*are they wearing matching uniforms ?*" shown in Figure 1, NeuralCoref takes the question Q4 and its history as inputs, and then generates a new question "*are the people wearing matching uniforms ?*" as a new Q4. As shown in Figure 3 (a), visual grounding model (Yang et al., 2019b) takes the i-th question Q_i and the image I as inputs and generates initial visual grounding features, as follows:

$$v_{g_i}^{(0)} = \text{VGM}(Q_i, I), \quad (4)$$

where $\text{VGM}(\cdot)$ denotes visual grounding model[2]. Then $v_{g_i}^{(0)}$ is sent to the multi-head self-attention

[1]Introduction and code of NeuralCoref are available at https://github.com/huggingface/neuralcoref. NeuralCoref is only used for visual grounding.

[2]Introduction and code are available at https://github.com/zyang-ur/onestage_grounding.

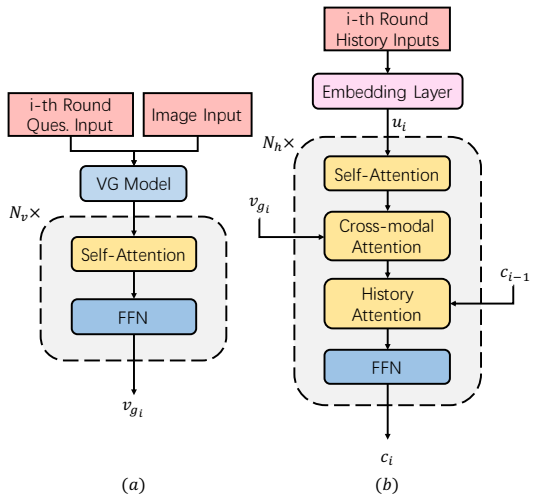

Figure 3: Framework of (a) Visual Grounding and (b) Multimodal Incremental Transformer Encoder (MITE).

layer followed by a position wise feed-forward network (FFN) layer (stacked N_v times) to generate the i-th visual grounding features as follows[3]:

$$\hat{v}_{g_i}^{n} = \text{MultiHead}\left(v_{g_i}^{(n-1)}, v_{g_i}^{(n-1)}, v_{g_i}^{(n-1)}\right), \quad (5)$$

where $n = 1, \ldots, N_v$ and $\text{MultiHead}(\cdot)$ denotes the multi-head self-attention layer (Vaswani et al., 2017), then

$$v_{g_i}^{(n)} = \text{FFN}\left(\hat{v}_{g_i}^{n}\right), \quad (6)$$

where $n = 1, \ldots, N_v$ and $\text{FFN}(\cdot)$ denotes the position wise feed-forward networks (Vaswani et al., 2017). After N_v layers computation, we obtain the final visual grounding features v_{g_i} by:

$$v_{g_i} = v_{g_i}^{(N_v)}, \quad (7)$$

Actually, there are some questions that do not contain any entities in the visual dialogue, such as "*anything else ?*". For such questions, we use the features of the whole image instead, i.e. $v_{g_i} = v$.

2.4 Multimodal Incremental Transformer

Inspired by the idea of incremental transformer (Li et al., 2019) which is originally designed for the single-modal dialogue task, we make an extension and propose a multimodal incremental transformer, which is composed of a Multimodal Incremental Transformer Encoder (MITE) and a Gated Cross-Attention Decoder (GCAD). The MITE uses an incremental encoding scheme to encode multi-turn dialogue history with an understanding of the image. The GCAD leverages the outputs from both the encoder and visual grounding via the gated cross-attention layer to fuse the two modal information in order to generate a contextually and visually coherent response word by word.

2.4.1 MITE

To effectively encode multi-turn utterances grounded in visual content, we design the Multimodal Incremental Transformer Encoder (MITE). As shown in Figure 3 (b), at the i-th round, where $i = 1, 2, ..., t-1$, the MITE takes the visual grounding features v_{g_i}, the dialogue history features u_i and the context state c_{i-1} as inputs, and utilizes attention mechanism to incrementally build up the representation of the relevant dialogue history and the associated image regions, and then outputs the new context state c_i. This process can be stated recursively as follows:

$$c_i = \text{MITE}\left(v_{g_i}, u_i, c_{i-1}\right), \quad (8)$$

where $\text{MITE}(\cdot)$ denotes the encoding function, c_i denotes the context state after the dialogue history features u_i and the visual grounding features v_{g_i} being encoded, and c_0 is the dialogue history features u_0.

As shown in Figure 3 (b), we use a stack of N_h identical layers to encode v_{g_i}, u_i and c_{i-1}, and to generate c_i. Each layer consists of four sub-layers. **The first sub-layer** is a multi-head self-attention for the dialogue history:

$$\text{A}^{(n)} = \text{MultiHead}\left(\text{C}^{(n-1)}, \text{C}^{(n-1)}, \text{C}^{(n-1)}\right), \quad (9)$$

where $n = 1, \ldots, N_h$, $\text{C}^{(n-1)}$ is the output of the last layer N_{n-1}, and $\text{C}^{(0)}$ is the dialog history features u_i. **The second sub-layer** is a multi-head cross-modal attention:

$$\text{B}^{(n)} = \text{MultiHead}\left(\text{A}^{n}, v_{g_i}, v_{g_i}\right), \quad (10)$$

where v_{g_i} is the visual grounding features. **The third sub-layer** is a multi-head history attention:

$$\text{F}^{(n)} = \text{MultiHead}\left(\text{B}^{(n)}, \text{c}_{i-1}, \text{c}_{i-1}\right), \quad (11)$$

where c_{i-1} is the context state after the previous dialogue history features u_{i-1} being encoded. That's why we call this encoder "Multimodal Incremental Transformer". **The fourth sub-layer** is a position wise feed-forward network (FFN):

$$\text{C}^{(n)} = \text{FFN}\left(\text{F}^{(n)}\right). \quad (12)$$

[3]For simplicity, we omit the descriptions of layer normalization and residual connection.

We use c_i to denote the final representation at N_h-th layer:

$$c_i = \mathrm{C}^{(N_h)}. \tag{13}$$

The mulitmodal incremental transformer encoder at the current turn t, i.e., the bottom one in Figure 2, has the same structure as all the other MITEs but takes the visual grounding features v_{g_t}, the current question features q_t and the context state c_{t-1} as inputs and generates the final context state c_t.

2.4.2 GCAD

Motivated by the real-world human cognitive process, we design a Gated Cross-Attention Decoder (GCAD) shown in Figure 2, which takes the masked answer features $a_{<z}$ (where $z = 1, 2, ..., Z$ and Z is the length of the answer), encoder outputs c_t and visual grounding features v_{g_t} as inputs, and generates contextually and visually coherent responses grounded in an image. GCAD is composed of a stack of N_y identical layers, each of which has three sub-layers.

The first sub-layer is a multi-head self-attention as follows:

$$\mathrm{J}^{(n)} = \mathrm{MultiHead}\left(\mathrm{R}^{(n-1)}, \mathrm{R}^{(n-1)}, \mathrm{R}^{(n-1)}\right), \tag{14}$$

where $n = 1, \ldots, N_y$, $\mathrm{R}^{(n-1)}$ is the output of the previous layer, and $\mathrm{R}^{(0)}$ is the masked answer features $a_{<z}$.

The second sub-layer is a multi-head gated cross-modal attention layer (GCA) as shown in Figure 4, calculated as:

$$\mathrm{P}^{(n)} = \alpha^{(n)} \circ \mathrm{E}^{(n)} + \beta^{(n)} \circ \mathrm{G}^{(n)}, \tag{15}$$

where $n = 1, \ldots, N_y$, $\circ$ denotes Hadamard product, $\mathrm{E}^{(n)}$ and $\mathrm{G}^{(n)}$ denote the outputs of two cross-attention functions, computed as follows:

$$\mathrm{E}^{(n)} = \mathrm{MultiHead}\left(\mathrm{J}^{(n)}, c_t, c_t\right), \tag{16}$$

$$\mathrm{G}^{(n)} = \mathrm{MultiHead}\left(\mathrm{J}^{(n)}, v_{g_t}, v_{g_t}\right), \tag{17}$$

where $\alpha^{(n)}$, $\beta^{(n)}$ are two gates[4]:

$$\alpha^{(n)} = \sigma\left(W_E[\mathrm{J}^{(n)}, E^{(n)}] + b_E\right), \tag{18}$$

$$\beta^{(n)} = \sigma\left(W_G[\mathrm{J}^{(n)}, G^{(n)}] + b_G\right), \tag{19}$$

where σ denotes sigmoid function, W_E, W_G, b_E, b_G are learnable parameters, and $[\cdot, \cdot]$ indicates concatenation.

[4]Our inspiration comes from Cornia et al. (2020).

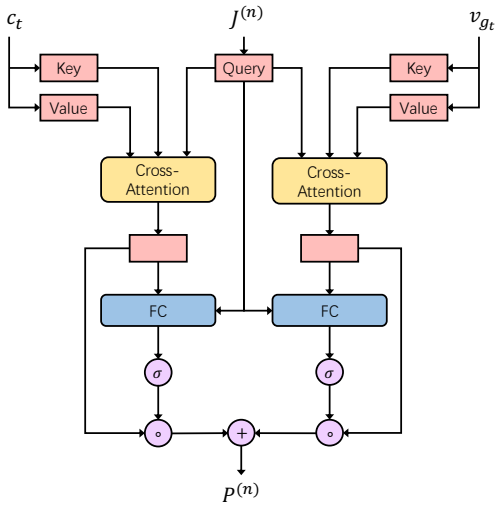

Figure 4: Framework of Gated Cross-Attention (GCA) in the Deocer.

The third sub-layer is a position wise feed-forward network (FFN):

$$\mathrm{R}^{(n)} = \mathrm{FFN}\left(\mathrm{P}^{(n)}\right). \tag{20}$$

We use r_z to denote the final representation at N_y-th layer:

$$r_z = \mathrm{R}^{(N_y)}. \tag{21}$$

Finally, we use softmax to get the word probabilities $\hat{a}_z$:

$$\hat{a}_z = \mathrm{softmax}(r_z). \tag{22}$$

3 Experiments

3.1 Datasets

We conduct experiments on the VisDial v0.9 and v1.0 datasets (Das et al., 2017) to verify our approach. VisDial v0.9 contains 83k dialogs on COCO-train (Lu et al., 2017) and 40k dialogs on COCO-val images as test set, for a total of 1.23M dialog question-answer pairs. VisDial v1.0 dateset is an extension of VisDial v0.9 dateset with additional 10k COCO-like images from Flickr. VisDial v1.0 dateset contains 123k, 2k and 8k images as train, validation and test splits, respectively.

3.2 Implementation and Evaluation

Implementation Details. Following previous work (Das et al., 2017), in order to represent words we firstly lowercase all the texts and convert digits to words, and then remove contractions before tokenization. The captions, questions and answers are further truncated to ensure that they are not longer than 40, 20 and 20 tokens, respectively. We construct the vocabulary of tokens that appear at least

Model	Object	Vis-G	MRR ↑	R@1 ↑	R@5↑	R@10 ↑	Mean ↓
AP (Das et al., 2017)	×	×	37.35	23.55	48.52	53.23	26.50
NN (Das et al., 2017)	×	×	42.74	33.13	50.83	58.69	19.62
LF (Das et al., 2017)	×	×	51.99	41.83	61.78	67.59	17.07
HREA (Das et al., 2017)	×	×	52.42	42.28	62.33	68.71	16.79
MN (Das et al., 2017)	×	×	52.59	42.29	62.85	68.88	17.06
HCIAE (Lu et al., 2017)	×	×	53.86	44.06	63.55	69.24	16.01
CorefNMN (Kottur et al., 2018)	×	×	53.50	43.66	63.54	69.93	15.69
CoAtt (Wu et al., 2018)	×	×	55.78	46.10	65.69	71.74	14.43
RvA (Niu et al., 2019)	✓	×	55.43	45.37	65.27	72.97	**10.71**
DVAN (Guo et al., 2019b)	✓	×	55.94	46.58	65.50	71.25	14.79
VDBERT (Wang et al., 2020)	✓	×	55.95	46.83	65.43	72.05	13.18
LTMI (Nguyen et al., 2020)†	✓	×	55.85	46.07	65.97	72.44	14.17
DMRM (Chen et al., 2020a)	✓	×	55.96	46.20	66.02	72.43	13.15
MITVG	✓	✓	**56.83**	**47.14**	**67.19**	**73.72**	11.95

Table 1: Performance on VisDial val v0.9 (Das et al., 2017). † indicates that we re-implement the model. "Object" and "Vis-G" denote if the model uses object-level image features and visual grounding, respectively. Underline denotes the highest score among baselines. Our MITVG exceeds previous work on most of the metrics and achieves comparable performance.

Model	Object	Vis-G	MRR ↑	R@1 ↑	R@5↑	R@10 ↑	Mean ↓	NDCG ↑
MN (Das et al., 2017)‡	✓	×	47.99	38.18	57.54	64.32	18.60	51.86
HCIAE (Lu et al., 2017)‡	✓	×	49.07	39.72	58.23	64.73	18.43	59.70
CoAtt (Wu et al., 2018)‡	✓	×	49.64	40.09	59.37	65.92	17.86	59.24
Primary (Guo et al., 2019a)	✓	×	49.01	38.54	59.82	66.94	16.60	-
ReDAN (Gan et al., 2019)	✓	×	50.02	40.27	59.93	66.78	17.40	60.47
DMRM (Chen et al., 2020a)	✓	×	50.16	40.15	60.02	67.21	15.19	-
LTMI (Nguyen et al., 2020)†	✓	×	50.38	40.30	60.72	68.44	15.73	61.61
DAM (Jiang et al., 2020b)	✓	×	50.51	40.53	60.84	67.94	16.65	60.93
KBGN (Jiang et al., 2020a)	✓	×	50.05	40.40	60.11	66.82	17.54	60.42
MITVG	✓	✓	**51.14**	**41.03**	**61.25**	**68.49**	**14.37**	61.47

Table 2: Performance on VisDial val v1.0 (Das et al., 2017). ‡ denotes that all the models are re-implemented by Gan et al. (2019). Our MITVG outperforms previous work and achieves comparable performance.

5 times in the training split. To represent image regions, we use Faster R-CNN (Ren et al., 2015b) with ResNet-101 (He et al., 2016) finetuned on the Visual Genome dataset (Krishna et al., 2017), thus obtaining a 2048-dimensional feature vector for each region. The layers of our encoder, decoder and visual grounding module are all set to 3. The number of attention heads in multi-head attention is 8 and the filter size is 2048. The word embedding is shared by the history, questions and responses. The dimension of word embedding is set to 512 empirically. We use Adam (Kingma and Ba, 2014) for optimization, following the learning rate scheduling strategy of Vaswani et al. (2017). Our model is implemented using PyTorch v1.0, Python v3.6, and provides out of the box support with CUDA 9 and CuDNN 7. We train our model on TITAN XP with 8 GPUs. For each epoch, we spend about 9,000 seconds on training the model. The total parameters are about 56.79M.

Before we train our model, we use three external tools for image features extracting, reference resolution and visual grounding.

Image Features Extracting We extract image features of VisDial images, using a Faster-RCNN (Ren et al., 2015b) with ResNet-101 (He et al., 2016) pre-trained on Visual Genome (Krishna et al., 2017), introduction and code from https://github.com/peteanderson80/bottom-up-attention.

Reference Resolution we use NeuralCoref v4.0 for reference resolution, which is developed by huggingface. Introduction and code are available at https://github.com/huggingface/neuralcoref.

Visual Grounding We use One-Stage Visual Grounding Model (Yang et al., 2019b) to obtain the visual grounding features. Introduction and code are available at https://github.com/zyang-ur/onestage_grounding.

Automatic Evaluation. We use a retrieval setting to evaluate individual responses at each round of a dialogue, following Das et al. (2017). Specif-

ically, at test time, apart from the image, ground truth dialogue history and the question, a list of 100-candidate answers is also given. The model is evaluated on retrieval metrics: (1) rank of human response (Mean, the lower the better), (2) existence of the human response in $top-k$ ranked responses, i.e., R@k (3) mean reciprocal rank (MRR) of the human response (the higher the better) and (4) normalized discounted cumulative gain (NDCG) for VisDial v1.0 (the higher the better). During evaluation, we use the log-likelihood scores to rank candidate answers.

Human Evaluation. We randomly extract 100 samples for human evaluation according to Wu et al. (2018), and then ask 3 human subjects to guess whether the last response in the dialogue is human-generated or machine-generated. If at least 2 of them agree it is generated by a human, we think it passes the Truing Test (M1). In addition, we record the percentage of responses that are evaluated better than or equal to human responses (M2), according to the human subjects' evaluation.

3.3 Main Results

We compare our proposed model to the state-of-the-art *generative models* developed in previous work. Current encoder-decoder based generative models can be divided into tree facets. (1) Fusion-based models: LF (Das et al., 2017) and HREA (Das et al., 2017) directly encode the multimodal inputs and decode the answer. (2) Attention-based models: HCIAE (Lu et al., 2017), CoAtt (Wu et al., 2018), Primary (Guo et al., 2019a), ReDAN (Gan et al., 2019), DVAN (Guo et al., 2019b) and DMRM (Chen et al., 2020a), DAM, LTMI, KBGN. (3) Visual co-reference resolution models: CorefNMN (Kottur et al., 2018), RvA (Niu et al., 2019). (4) The pretraining model: VDBERT (Wang et al., 2020).

As shown in Table 1 and Table 2, our MITVG, which explicitly locates related objects guided by the textual entities and implements a multimodal incremental transformer to incrementally build the representation of the dialogue history and the image, achieves comparable performance on the VisDial v0.9 and v1.0 datasets. Specifically, our model outperforms previous work by a significant margin both on the VisDial v0.9 dataset (0.87 on MRR, 0.31 on R@1, 1.17 on R@5, 0.75 on R10) and the VisDial v1.0 dataset (0.98 on MRR, 0.76 on R@1, 1.23 on R@5, 1.28 on R10, 0.82 on Mean, and 1.00 on NDCG). The improvement of R@10 is the largest and our method also gains a large increase on MRR and R@1 due to the explicit modeling of multiple modalities (Seeing Sec 3.5 for further quantitative analysis).

	DMRM	MITVG
Method 1 (M1)	0.62	**0.76**
Method 2 (M2)	0.59	**0.70**

Table 3: Human evaluation on 100 sampled responses on VisDial val v1.0. M1: percentage of responses pass the Turing Test. M2: percentage of responses evaluated better than or equal to human responses.

As shown in Table 3, we conduct human study to further prove the effectiveness of our model. Our model achieves the highest scores both on the metric M1 (0.76) and M2 (0.70) compared with the previous model, DMRM (Chen et al., 2020a). These results show that our model can generate a better contextually and visually coherent response.

3.4 Ablation Study

We also conduct an ablation study to illustrate the validity of our proposed Multimodal Incremental Transformer with Visual Grounding. The results are shown in Table 4.

We implement Multimodal Incremental Transformer without Visual Grounding ('MITVG w/o VG') to verify the validity of visual grounding. As shown in Table 4, comparing 'MITVG w/o VG' with MITVG, we find the metrics decrease obviously (0.46 on MRR, 0.60 on R@1, 0.68 on R@5, 0.46 on R@10 and 0.59 on Mean) if visual grounding is deleted from MITVG. This observation demonstrates the validity of visual grounding.

To verify the effectiveness of the incremental transformer architecture, we implement a Multimodal Incremental LSTM without Visual Grounding ('MI-LSTM w/o VG'). A 3-layer bidirectional LSTM (Schuster and Paliwal, 1997) with multi-head attention and a 1-layer LSTM with GCA are applied for encoder and decoder, respectively. All the LSTM hidden state size is 512. Results in Table 4 demonstrate the effectiveness of our incremental transformer architecture (compare 'MITVG w/o VG' with 'MI-LSTM w/o VG'). Results from the comparison between 'MITVG w/o VG' and DMRM (Chen et al., 2020a) also show the validity of our incremental transformer to some extent.

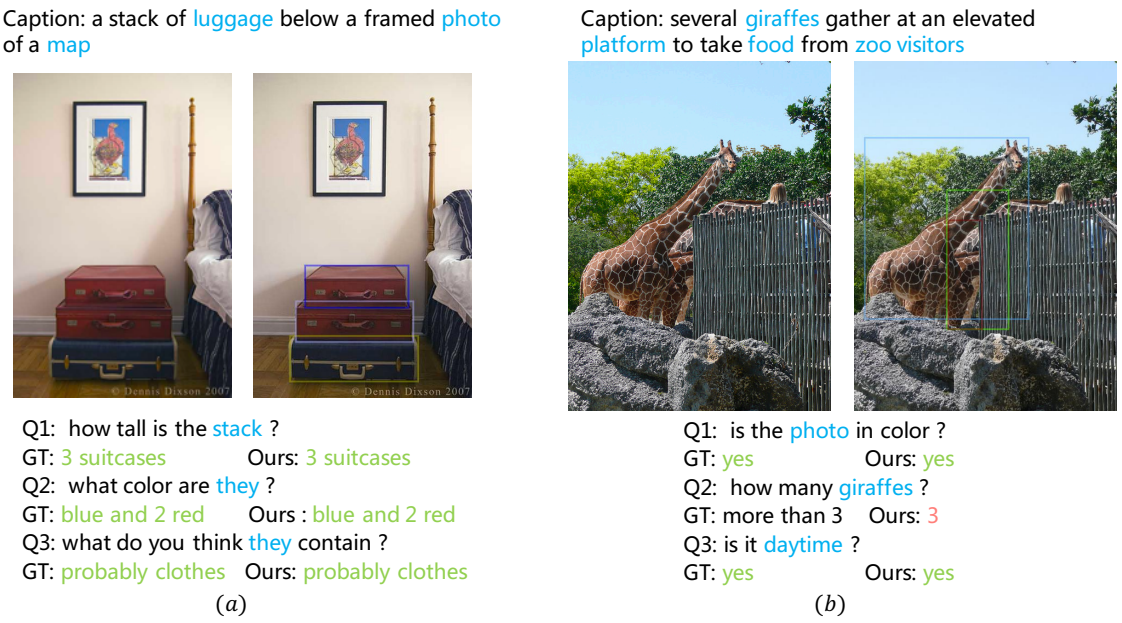

Figure 5: Case study. The text marked in blue indicates the dialogue topic. The answers marked in green and red indicate the right and wrong answers, respectively. Our MITVG often generates right responses (marked in green) in keeping with human answers.

Model	MRR	R@1	R@5	R@10	Mean
DMRM	50.16	40.15	60.02	67.21	15.19
MITVG	**51.14**	**41.03**	**61.25**	**68.49**	**14.37**
MITVG w/o VG	50.68	40.43	60.57	68.03	14.96
MI-LSTM w/o VG	50.02	39.85	59.86	67.16	15.78

Table 4: Ablation study of our proposed model on VisDial val v1.0. "MI-LISM" indicates Multimodal Incremental LSTM. "VG" indicates visual grounding.

	Train	Validation	Test
VisDial v0.9	2.04	1.95	-
VisDial v1.0	2.05	1.93	1.93

Table 5: Average number of the grounded objects in each question.

3.5 Case Study

As shown in Table 5, we calculate the average number of the objects associated with entities in each question for assistant analysis. As shown in Figure 5 (a), owing to the explicit understanding of visual content via visual grounding and the multimodal incremental transformer architecture, our MITVG generates responses in keeping with human answers. For example, while answering the question Q1 "*how tall is the stack ?*" and Q2 "*what color are they ?*", our model grounds the three suitcases accurately via visual grounding, thus giving the accurate responses "*3 suitcases*" and "*blue and 2 red*". However, as shown in Figure 5 (b), for questions Q2, MITVG gives a wrong answer because it focuses on wrong number of objects in the question by visual grounding.

4 Related Work

Visual Dialogue. Our work touches two branches of the research in visual dialogue. One is how to leverage image features. Niu et al. (2019) utilize object-level image features as visual attention and refine it by recursively reviewing the dialog history. Gan et al. (2019) and Chen et al. (2020a) regard the object-level image features as visual memory to infer answers progressively through multiple steps. The other is how to model dialogue history. Yang et al. (2019a) propose a new training paradigm inspired by actor-critic policy gradient (Sutton et al., 1999) for history-advantage training. Guo et al. (2020) represent each turn dialogue history with visual content as a node in a context-aware graph neural network. Park et al. (2020) refine history information from both topic aggregation and context matching. Different from these approaches, we explicitly establish specific mapping of objects and textual entities to exclude undesired visual content via visual grounding, and model multi-turn structure of the dialogue based on visual grounding to develop a unified representation combining multi-turn utterances

along with the relevant objects.

Incremental Structures. There are some successes on introducing the incremental structure into tasks related to dialog systems (Zilka and Jurcicek, 2015; Coman et al., 2019; Li et al., 2019; Das et al., 2017). In particular, Coman et al. (2019) propose an incremental dialog state tracker which is updated on a token basis from incremental transcriptions. Li et al. (2019) devise an incremental transformer to encode multi-turn utterances along with knowledge in related documents for document grounded conversations. Das et al. (2017) propose a dialog-RNN to produce an encoding for this round and a state for next round. Our model is different from these approaches mainly in two aspects: 1) we explicitly model the relationship between modalities, i.e., textual utterance and image objects, in visual dialogue through visual grounding; 2) based on the explicit association between modalities, our model incrementally encodes the dialogue history and the image with well-designed incremental multimodal architecture to sufficiently understand the dialogue content, thus generating better responses.

5 Conclusion

We propose a novel Multimodal Incremental Transformer with Visual Grounding for visual dialogue, named MITVG, which consists of two key parts: visual grounding and multimodal incremental transformer. Visual grounding aims to explicitly model the relationship between multiple modalities. Based on visual grounding, multimodal incremental transformer aims to explicitly model multi-turn dialogue history in the order of the dialogue. Experiments on the VisDial v0.9 and v1.0 datasets show that our model achieves comparable performance.

References

Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. 2020. History for visual dialog: Do we really need it? *arXiv preprint arXiv:2005.07493*.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic propositional image caption evaluation. *Adaptive Behavior*, 11(4):382–398.

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.

Feilong Chen, Fandong Meng, Jiaming Xu, Peng Li, Bo Xu, and Jie Zhou. 2020a. DMRM: A dual-channel multi-hop reasoning model for visual dialog. *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Xiuyi Chen, Fandong Meng, Peng Li, Feilong Chen, Shuang Xu, Bo Xu, and Jie Zhou. 2020b. Bridging the gap between prior and posterior knowledge selection for knowledge-grounded dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3426–3437.

Xiuyi Chen, Jiaming Xu, and Bo Xu. 2019. A working memory model for task-oriented dialog response generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2687–2693.

Andrei C Coman, Koichiro Yoshino, Yukitoshi Murase, Satoshi Nakamura, and Giuseppe Riccardi. 2019. An incremental turn-taking model for task-oriented dialog systems. *arXiv preprint arXiv:1905.11806*.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.

Zhe Gan, Yu Cheng, Ahmed EI Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. 2019. Multi-step reasoning via recurrent dual attention for visual dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6463–6474.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in Neural Information Processing Systems*, pages 2296–2304.

Dalu Guo, Chang Xu, and Dacheng Tao. 2019a. Image-question-answer synergistic network for visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10434–10443.

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

Dan Guo, Hui Wang, Hanwang Zhang, Zheng-Jun Zha, and Meng Wang. 2020. Iterative context-aware graph inference for visual dialog. *arXiv preprint arXiv:2004.02194*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

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. *Proceedings of the 28th ACM International Conference on Multimedia*.

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. *arXiv preprint arXiv:2007.03310*.

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

Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2018. Visual coreference resolution in visual dialog using neural module networks. *ArXiv*, abs/1809.01816.

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

Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. 2019. Incremental transformer with deliberation decoder for document grounded conversations. *arXiv preprint arXiv:1907.08854*.

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. In *Advances in Neural Information Processing Systems*, pages 314–324.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297.

Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. 2020. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. *Proceedings of the European Conference on Computer Vision*.

Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. 2020. Efficient attention mechanism for visual dialog that can handle all the interactions between multiple inputs. *Proceedings of the European Conference on Computer Vision*.

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*.

Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. 2020. Two causal principles for improving visual dialog. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015a. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems*, pages 2953–2961.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015b. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Richard S Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12, pages 1057–1063.

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

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*.

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.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of International Conference on Machine Learning*, pages 2048–2057.

Tianhao Yang, Zheng-Jun Zha, and Hanwang Zhang. 2019a. Making history matter: History-advantage sequence training for visual dialog. In *The IEEE International Conference on Computer Vision (ICCV)*.

Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. 2019b. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4683–4693.

Lukas Zilka and Filip Jurcicek. 2015. Incremental lstm-based dialog state tracker. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (Asru)*, pages 757–762. IEEE.