

# DualVD: An Adaptive Dual Encoding Model for Deep Visual Understanding in Visual Dialogue

Xiaoze Jiang<sup>1,2</sup> Jing Yu<sup>1\*</sup> Zengchang Qin<sup>2\*</sup> Yingying Zhuang<sup>1,2</sup> Xingxing Zhang<sup>3</sup> Yue Hu<sup>1</sup> Qi Wu<sup>4</sup>

<sup>1</sup> Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

<sup>2</sup> Intelligent Computing and Machine Learning Lab, School of ASEE, Beihang University, Beijing, China

<sup>3</sup> Microsoft Research Asia, Beijing, China

<sup>4</sup> University of Adelaide, Australia

{yujing02, huyue}@iie.ac.cn, {xzjiang,zcqin}@buaa.edu.cn, xizhang@microsoft.com, qi.wu01@adelaide.edu.au

## Abstract

Different from Visual Question Answering task that requires to answer only one question about an image, Visual Dialogue involves multiple questions which cover a broad range of visual content that could be related to any objects, relationships or semantics. The key challenge in Visual Dialogue task is thus to learn a more comprehensive and semantic-rich image representation which may have adaptive attentions on the image for variant questions. In this research, we propose a novel model to depict an image from both visual and semantic perspectives. Specifically, the visual view helps capture the appearance-level information, including objects and their relationships, while the semantic view enables the agent to understand high-level visual semantics from the whole image to the local regions. Furthermore, on top of such multi-view image features, we propose a feature selection framework which is able to adaptively capture question-relevant information hierarchically in fine-grained level. The proposed method achieved state-of-the-art results on benchmark Visual Dialogue datasets. More importantly, we can tell which modality (visual or semantic) has more contribution in answering the current question by visualizing the gate values. It gives us insights in understanding of human cognition in Visual Dialogue.

## Introduction

To understand the real world by analyzing vision and language together is a priority for AI to achieve human-like abilities, which enables the development of diverse applications, such as Visual Question Answering (VQA) (Agrawal et al. 2017), Referring Expressions (Wang et al. 2019), Image Captioning (Johnson, Karpathy, and Fei-Fei 2016), etc. To move a step further, this work focuses on the Visual Dialogue (Das et al. 2017) problem, which requires the agent to answer a series of questions in natural language regarding an image. It is more challenging because it demands the agent to adaptively focus on diverse visual content with respect to the current question, while other vision-language problems mostly attend to some specific objects or regions. Considering the dialogue in Figure 1: Given “*Q1: Is the man on the*

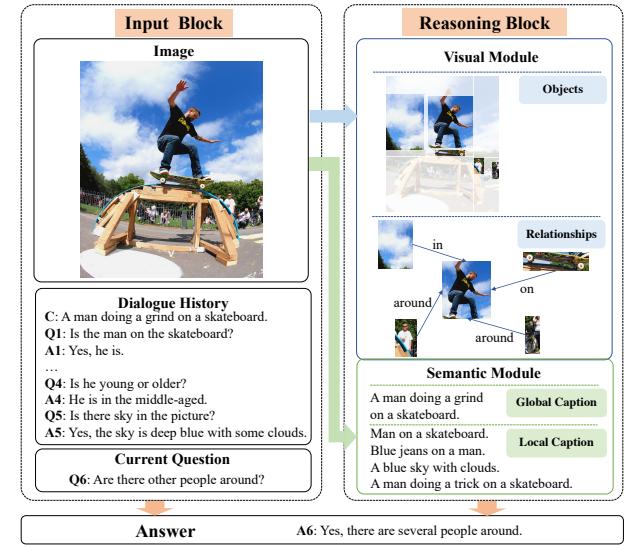


Figure 1: An illustration of DualVD. Left: the input of the dialogue system. Right: visual and semantic modules designed to adaptively understand the visual content like humans. The answer is inferred depending on multi-modal evidence.

*skateboard?*”, the agent should be aware of the foreground visual content, i.e. *the man, the skateboard*, while “*Q5: Is there sky in the picture?*” changes the attention of the agent to the background of *sky*. Besides appearance-level questions like *Q1* and *Q5*, “*Q4: Is he young or older?*” requires the agent to reason about the visual content for higher-level semantics. How to adaptively capture the desired visual content through dialogue becomes one of the most critical challenges in visual dialogue.

The typical solution for visual dialogue is to firstly fuse visual (*i.e.* image) features and textual (*i.e.* dialogue history, current question) features together and then to infer the correct answer. Most approaches focus on enhancing the textual representations by recovering the dialogue relational structure (Zheng et al. 2019), imperfect dialogue history (Yang, Zha, and Zhang 2019), and dialogue consistency (Qi et al. 2018). However, the role of visual information is at

\*Corresponding authors: Jing Yu and Zengchang Qin.  
Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

present less studied. Existing models simply use CNN (Simonyan and Zisserman 2014) or R-CNN (Ren et al. 2017) to extract visual features and focus on the question-relevant content. Such visual features have limited expressive ability due to the monolithic representations (Wang et al. 2019). On one hand, questions in a visual dialogue refer to a wide range of visual content, including objects, relationships and high-level semantics, which can not be covered by monolithic features. On the other hand, the referred visual content may change remarkably from visual appearance to high-level semantics through the dialogue, which is difficult for monolithic features to capture.

Our work is inspired by the Dual-coding theory (Paivio 1971) of human cognition process. *Dual-coding theory* postulates that our brain encodes information in two ways: *visual imagery* and *textual associations*. When asked to act upon a concept, our brain retrieves either images or words, or both simultaneously. The ability to encode a concept by two different ways strengthens the capacity of memory and understanding. Inspired by the cognitive process, we first propose a novel scheme to comprehensively depict an image from both visual and semantic perspectives, where the major objects and their relationships are kept in the visual view while the higher-level abstraction is provided in the semantic view. We propose a model called *Dual Encoding Visual Dialogue (DualVD)* to adaptively select question-relevant information from the image in a hierarchical mode: intra-modal selection first captures the visual and semantic information individually from the object-relational visual features and global-local semantic features; then inter-modal selection obtains the joint visual-semantic knowledge by correlating vision and semantics. This hierarchical framework imitates human cognition process to capture targeted visual clues from multiple perceptual views and semantic levels.

The main contributions are summarized as follows: (1) We exploit the possibility of cognition in visual dialogue by depicting an image from both visual and semantic views, which covers a broad range of visual content referred by most of questions in the visual dialogue task; (2) We propose a hierarchical visual information selection model, which is able to progressively select question-adaptive clues from intra-modal and inter-modal information for answering diverse questions. It supports explicit visualization in visual-semantic knowledge selection and reveals which modality has more contribution to answer the question; (3) The proposed model outperforms state-of-the-art approaches on benchmark visual dialogue datasets, which demonstrates the feasibility and effectiveness of the proposed model. The code is available at <https://github.com/JXZe/DualVD>.

## Related Work

**Visual Question Answering (VQA)** focuses on answering arbitrary natural language questions conditioned on an image. The typical solutions in VQA build multi-modal representations upon CNN-RNN architecture (Ren, Kiros, and Zemel 2015; Qi et al. 2017). Existing approaches incorporate context-aware visual features. For example, (Ren, Kiros, and Zemel 2015) applies CNN features of the whole image as global context, (Xu and Saenko 2016; Anderson et

al. 2018) adopt patches and salient objects learned by attention mechanism as the region context, and (Gao et al. 2018; Li et al. 2019b) exploits inter-object relationships via graph attention networks or convolutional networks to model the relational context. However, how to leverage the external visual-semantic knowledge to learn more informative relational representations for better semantic understanding has not been well exploited yet. Another emerging line of work represents visual content explicitly by natural language and solves VQA as a reading comprehension problem. In (Li et al. 2019a), the image is wholly converted into descriptive captions, which preserves information at semantic-level in textual domain. However, this kind of approaches use the generated captions, which could not be correct as we desired, and that they fully abandon the informative and subtle visual features. Besides the specific tasks, our model has notable progress compared to the above approaches. We adopt dual encoding mechanism to provide both appearance-level and semantic-level visual information, so that it incorporates the strong points of the above two kinds of approaches.

**Visual Dialogue** aims to answer a current question conditioned on an image and dialogue history. Most existing works are based on late fusion framework and focused on modeling the dialogue history. Sequential co-attention mechanism (Qi et al. 2018) enables the model to identify question-relevant image regions and dialogue history to keep the dialogue consistency. (Yang, Zha, and Zhang 2019) introduces false response in dialogue history for an adverse critic on the historic error. (Zheng et al. 2019) introduces an Expectation Maximization algorithm to infer the dialogue structure and the answers via graph neural networks. By contrast to extensive study on modeling dialogue history, the image content has been less studied. Although some works devise attention mechanism to focus on the essential visual features most relevant to the question and dialogue history, such monolithic visual representations still have limited expressive abilities. In this work, we exploit the role of visual information in visual dialogue. Different from existing works merely modeling the appearance, our model is able to adaptively capture visual and semantic information in a hierarchical mode inspired by the Dual-coding theory of human cognition process to provide adequate visual clues for diverse questions in visual dialogue.

## Methodology

The visual dialogue task can be described as follows: given an image  $I$  and its caption  $C$ , a dialogue history till round  $t-1$ ,  $H_t = \{C, (Q_1, A_1), \dots, (Q_{t-1}, A_{t-1})\}$ , and the current question  $Q_t$ , the task is to rank a list of 100 candidate answers  $\mathbb{A} = \{A_1, A_2, \dots, A_{100}\}$  and return the best answer  $A_t$  to  $Q_t$ . In this section, we first introduce the idea of depicting an image from both visual and semantic perspectives. It covers a broad range of visual content like objects, relationships, global semantics and local semantics. Then we introduce a hierarchical feature selection approach to adaptively capture question-relevant visual-semantic information. Our model is based on the late fusion (LF) framework (Das et al. 2017), which will be described at the end of this section.

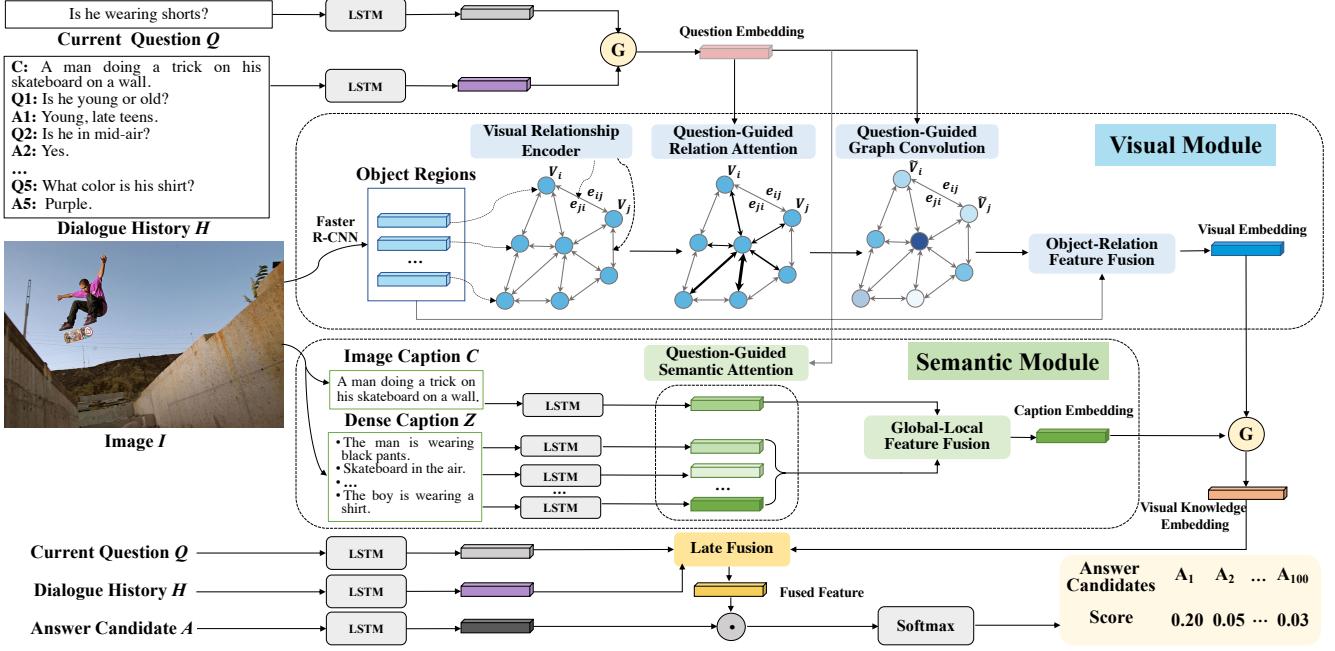


Figure 2: Overview structure of the DualVD model for visual dialogue. The model mainly contains two parts: Visual Module and Semantic Module, where “G” represents gate operation given inputs.

## Visual-Semantic Dual Encoding

In visual dialogue, two types of information play the primary role to depict an image and answer the diverse questions: visual information and semantic information (Figure 2). For visual information, the major objects and relationships should be kept. In semantic information, higher-level abstraction of the image content should be provided, which involves prior knowledge and complex cognition. In this section, we introduce a dual encoding scheme to generate both visual and semantic representations to depict an image. A scene graph is proposed to represent the visual information while multi-level captions in natural language are leveraged to represent the semantic information. These representations are served as the input of our DualVD model.

**Scene Graph Construction** Each image is represented as a scene graph. Let  $V = \{v_i\}^N$  denotes its nodes, which represents objects detected by a pre-trained object detector and let  $E = \{e_{ij}\}^{N \times N}$  denotes its edges, which represents the semantic visual relationships embedded by our visual relationship encoder. We use a pre-trained Faster-RCNN (Ren et al. 2017) to detect  $N$  objects in an image and describe the object  $v_i$  as a 2,048-dimensional vector, denoted by  $h_i$ . The visual relationship encoder (Zhang et al. 2019), which is pre-trained on a visual relationship benchmark, i.e. *GQA* (Hudson and Manning 2019), encodes relationships between the subject  $v_i$  and object  $v_j$  as a 512-dimensional relation embedding, denoted as  $r_{ij}$ . We assume that certain relationship exists between any pair of objects by considering “unknown-relationship” as a special kind of relationship. Therefore, the scene graph we constructed is fully-connected.

The visual relationship encoder embeds the relationships

between objects into a semantic space which is aligned with their corresponding descriptions in natural language. Such continuous representations instead of discrete labels can preserve the discriminative capability and contextual awareness. Inspired by recent work (Zhang et al. 2019), our encoder consists of a visual part and a textual part. The visual part takes three CNN feature maps corresponding to the visual regions of subject, object and their union region as input and outputs the three encoded embeddings  $x^s$ ,  $x^o$  and  $x^r$ . The textual part uses a shared GRU to encode the annotations and yield textual embeddings. The loss function is designed to minimize the cosine similarity between the embeddings of positive visual-textual pairs and alienate negative pairs. The union embedding  $x^r$  is served as the visual relationship representation  $r_{ij}$  between  $v_i$  and  $v_j$ .

**Multi-level Image Captions** The advantages of captions compared to visual features lie in that captions are represented by natural language with high-level semantics, which can provide straightforward clues for the questions without “heterogeneous gap”. Global image caption  $C$  (provided by the dataset) is beneficial to response to questions exploring the scene. Meanwhile, dense captions (Johnson, Karpathy, and Fei-Fei 2016), denoted as  $Z = \{z_1, z_2, \dots, z_k\}$  ( $k$  is the number of dense captions), provide a set of local-level semantics, including the object properties (position, color, shape, etc.), the prior knowledge related to the objects (weather, species, emotion, etc.), and the relationships between objects (interactions, spatial positions, comparison, etc.). The words in both  $C$  and  $Z$  are represented by concatenated GloVe (Pennington, Socher, and Manning 2014) and ELMo (Peters et al. 2018) word embeddings. Then  $C$

and  $Z$  are separately encoded with two different LSTMs, denoted as  $\tilde{C}$  and  $\tilde{Z} = \{\tilde{z}_1, \tilde{z}_2, \dots, \tilde{z}_k\}$ , respectively.

## Adaptive Visual-Semantic Knowledge Selection

On top of the visual and semantic image representations, we propose a novel feature selection framework to adaptively select question-relevant information from the image. Under the guidance of the current question, the feature selection process is devised in a hierarchical mode: intra-modal selection first captures the visual and semantic information respectively from the *visual module* and *semantic module*; then inter-modal selection obtains the desired visual knowledge from both the visual module and semantic module via *selective visual-semantic fusion*. The advantages of such hierarchical framework is that it can explicitly reveal the progressive feature selection mode and preserve fine-grained information as much as possible.

**Visual Module** This module is presented on the top of Figure 2. Based on the constructed scene graph introduced in *Scene Graph Construction*, we aim to select question-relevant relation information and object information. For relation information, we propose a relation-based graph attention network to enrich the object representations with question-aware relationships. It mainly consists of two units: *Question-Guided Relation Attention* highlights the critical relationships and *Question-Guided Graph Convolution* enriches the object features by its relation-critical neighbors. For object information, we highlight the most informative objects to answer the question. Finally, the clues of objects and relationships are further fused in *Object-Relation Information Fusion* to obtain the question-relevant visual content.

*Question-Guided Relation Attention*: The question-guided relation attention examines all the relationships to highlight the ones most relevant to the question. First, we select question-relevant information from the dialogue history to merge into the question representation via a gate operation, which is defined as:

$$gate_t^q = \sigma(\mathbf{W}_q[\tilde{H}_t, \tilde{Q}_t] + b_q) \quad (1)$$

$$\tilde{Q}_t^g = \mathbf{W}_1(gate_t^q \circ [\tilde{H}_t, \tilde{Q}_t]) + b_1 \quad (2)$$

where “[., .]” denotes concatenation, “ $\circ$ ” denotes the element-wise product. Each word is represented by concatenating the hidden states extracted from pre-trained GloVe and ELMo models. Then dialogue history  $H_t$  and the current question  $Q_t$  are separately encoded with two different LSTMs, denoted as  $\tilde{H}_t$  and  $\tilde{Q}_t$ , respectively.  $gate_t^q$  is a vector of gate values over  $\tilde{H}_t$  and  $\tilde{Q}_t$ ,  $\mathbf{W}_1$  (as well as  $\mathbf{W}_2, \dots, \mathbf{W}_7$  mentioned below) is the linear transformation layer and  $\tilde{Q}_t^g$  is the encoded history-aware question features.

The attention weights  $\alpha_{ij}$  of all the visual relationships are calculated under the guidance of the question  $\tilde{Q}_t^g$ :

$$\alpha_{ij} = softmax(\mathbf{W}_\rho(\mathbf{W}_2\tilde{Q}_t^g \circ \mathbf{W}_3r_{ij}) + b_r) \quad (3)$$

Each relation embedding is updated based on the attention importance. Formally defined as:

$$\tilde{r}_{ij} = \alpha_{ij}r_{ij} \quad (4)$$

where  $\tilde{r}_{ij}$  is the question-guided relation embedding.

*Question-Guided Graph Convolution*: This module further updates each object’s representation under the guidance of questions by aggregating information from its neighborhood and the corresponding relationships. Given the feature  $h_j$  of object  $v_j$  and its relation embedding  $\tilde{r}_{ij}$ , the attention value of  $v_j$  w.r.t.  $v_i$  is calculated as:

$$\beta_{ij} = softmax(\mathbf{W}_g(\tilde{Q}_t^g \circ (\mathbf{W}_4[h_j, \tilde{r}_{ij}])) + b_g) \quad (5)$$

The obtained attention values for all the neighbors of  $v_i$  are used to compute a linear combination of their features, which serves as the updated representation  $\tilde{h}_i$  for  $v_i$ :

$$\tilde{h}_i = \sum_{j=1}^N \beta_{ij} h_j \quad (6)$$

Since the scene graph is a fully connected graph, the number of neighbors  $N$  for each object is equal to the number of objects detected in each image.

*Object-Relation Information Fusion*: In visual dialogue, the object appearance and the visual relationships will contribute to infer the answer, but with different contributions. In this module, we adaptively fuse question-relevant object features from both original object feature  $h_i$  and relation-aware object feature  $\tilde{h}_i$  again by a gate, which is defined by:

$$gate_i^v = \sigma(\mathbf{W}_v[h_i, \tilde{h}_i] + b_v) \quad (7)$$

$$\tilde{h}_i^g = \mathbf{W}_5(gate_i^v \circ [h_i, \tilde{h}_i]) + b_5 \quad (8)$$

where  $\tilde{h}_i^g$  is the updated representation of object  $v_i$ . The whole image representation  $\tilde{I}$  is obtained as the weighted sum of the object representations. In order to strengthen the influence of the current question  $Q_t$  and the original object features on the retrieved visual clues, we calculate the attention value  $\gamma_i^v$  for  $h_i$  under the guidance of  $Q_t$ :

$$\gamma_i^v = softmax(\mathbf{W}_s(Q_t \circ (\mathbf{W}_6h_i)) + b_s) \quad (9)$$

Then the the whole representation of the image  $\tilde{I}$  can be updated by:

$$\tilde{I} = \sum_{i=1}^N \gamma_i^v \tilde{h}_i^g \quad (10)$$

**Semantic Module** This module aims to select and merge question-relevant semantic information from global and local captions with a *Question-Guided Semantic Attention* module and a *Global-Local Information Fusion* module. The semantic module is located in the middle of Figure 2.

*Question-Guided Semantic Attention*: The semantic attention mechanism highlights relevant captions at both global-level and local-level. This type of attention is guided by the current question which is enhanced with corresponding information from the dialogue history (as introduced above). According to the attention distribution, we enrich the caption representations in order to better adapt to the question. The attention value for each caption in  $m_i \in \{\tilde{C}, \tilde{z}_1, \tilde{z}_2, \dots, \tilde{z}_k\}$  is calculated as follows:

$$\delta_i^q = softmax((\mathbf{W}_{z1}\tilde{Q}_t^g + b_{z1})^T(\mathbf{W}_{z2}m_i + b_{z2})) \quad (11)$$

The caption representation for  $\tilde{C}$  and  $\tilde{Z}$  will be updated to  $\tilde{C}^q$  and  $\tilde{Z}^q$ :

$$\tilde{C}^q = \delta_1^q C \quad (12)$$

$$\tilde{Z}^q = \sum_{i=2}^{k+1} \delta_i^q \tilde{z}_{i-1} \quad (13)$$

*Global-Local Information Fusion:* Some questions are global-related while others are local-related. This step adaptively selects the information from the global caption  $\tilde{C}^q$  and local caption  $\tilde{Z}^q$  via a gate as described above:

$$gate^c = \sigma(\mathbf{W}_c[\tilde{C}^q, \tilde{Z}^q] + b_c) \quad (14)$$

$$\tilde{T} = \mathbf{W}_7(gate^c \circ [\tilde{C}^q, \tilde{Z}^q]) + b_7 \quad (15)$$

where  $\tilde{T}$  is the textural representations for the abstract visual semantics.

**Selective Visual-Semantic Fusion** When asked to answer a question, the agent will retrieve either the visual information or the semantic information individually, or both simultaneously. In this module, we design a gate operation to decide the contributions of the two modalities on the answer prediction. The gate operation and the final visual knowledge representation  $S$  are calculated as:

$$gate^s = \sigma(\mathbf{W}_s[\tilde{I}, \tilde{T}] + b_s) \quad (16)$$

$$S = gate^s \circ [\tilde{I}, \tilde{T}] \quad (17)$$

## Late Fusion and Discriminative Decoder

The full model consists of late fusion encoder and discriminative (softmax) decoder. The encoder first embeds each part in a dialogue tuple  $D = \{I, H_t, Q_t\}$ . Then we concatenate  $\tilde{H}_t$  and  $\tilde{Q}_t$  with the visual knowledge representation  $S$  into a joint input embedding for answer prediction. The decoder ranks all the answers from a set of 100 candidates  $\mathbb{A}$ . It first encodes each candidate via a common LSTM. Then a dot product followed by softmax operation is calculated between the joint input embedding and candidates to get the posterior probability over each candidate. We obtain the correct answer by ranking the candidates based on their posterior probabilities. Our model can also be applied to more complex decoders and fusion strategies, such as memory network, co-attention, adversarial network, etc. In this paper, we utilize the simple late fusion and discriminative decoder to highlight the advantages of our visual encoder.

## Experiments

**Datasets:** We conduct extensive experiments on datasets (Das et al. 2017): VisDial v0.9 and VisDial v1.0. For both datasets, the examples are split into “train”, “val” and “test” and each dialogue contains 10 rounds of question-answer pairs. VisDial v1.0 is an upgraded version of VisDial v0.9. For VisDial v0.9, all the splits are built on MSCOCO images. For VisDial v1.0, all the splits of VisDial v0.9 serve as “train” (120k), while “val” (2k) and “test” (8k) consist of dialogues on extra 10k COCO-like images from Flickr.

Table 1: Comparison on validation split of VisDial v0.9.

Model	MRR	R@1	R@5	R@10	Mean
LF	58.07	43.82	74.68	84.07	5.78
HRE	58.46	44.67	74.50	84.22	5.72
MN	59.65	45.55	76.22	85.37	5.46
SAN-QI	57.64	43.44	74.26	83.72	5.88
HieCoAtt-QI	57.88	43.51	74.49	83.96	5.84
AMEM	61.60	47.74	78.04	86.84	4.99
HCIAE	62.22	48.48	78.75	87.59	4.81
SF	62.42	48.55	78.96	87.75	4.70
CoAtt	63.98	50.29	80.71	88.81	4.47
CorefMN	<b>64.10</b>	<b>50.92</b>	80.18	88.81	4.45
VGNN	62.85	48.95	79.65	88.36	4.57
<b>DualVD</b>	62.94	48.64	<b>80.89</b>	<b>89.94</b>	<b>4.17</b>

**Evaluation Metrics:** We follow the metrics in (Das et al. 2017) to evaluate the response performance. In the test stage, the model is asked to rank 100 candidate answer options and evaluated by Mean Reciprocal Rank (MRR), Recall@ $k$  ( $k = 1, 5, 10$ ) and Mean Rank of human response (Mean) on both datasets. For VisDial v1.0, Normalized Discounted Cumulative Gain (NDCG) is added as an extra metric for more comprehensive analysis. Lower value for Mean and higher value for other metrics are desired.

**Implementation Details:** For the textual part, the maximum sentence length of the dialogue history, dense captions and the current question is all set to 20. The hidden state size of all the LSTM blocks is set to 512. We use Faster-RCNN with the ResNet-101 to detect object regions and extract the 2048-dimensional region features. Since some captions with low confidence are likely to introduce unexpected noise and too many captions will decrease the computation efficiency, we select the top 6 (the mean value of the caption distribution) dense captions in our model. We train all of our models by Adam optimizer with 16 epochs, where the mini-batch size is 15 and the dropout ratio is 0.5. For the strategy of learning rate, we first apply warm up strategy for 2 epoches with initial learning rate  $1 \times 10^{-3}$  and warm-up factor 0.2. Then we adopt cosine annealing learning strategy with initial learning rate  $\eta_{max}=1 \times 10^{-3}$  and termination learning rate  $\eta_{min}=3.4 \times 10^{-4}$  for the rest epoches.

## Overall Results

In Table 1 and Table 2, we compare DualVD with state-of-the-art discriminative models, namely LF (Das et al. 2017), HRE (Das et al. 2017), MN (Das et al. 2017), SAN-QI (Yang et al. 2016), HieCoAtt-QI (Lu et al. 2016), AMEM (Seo et al. 2017), HCIAE (Lu et al. 2017), SF (Jain, Lazebnik, and Schwing 2018), CoAtt (Qi et al. 2018), CorefMN (Kottur et al. 2018), VGNN (Zheng et al. 2019), LF-Att (Das et al. 2017), MN-Att (Das et al. 2017), RvA(Niu et al. 2019) and DL-61(Guo, Xu, and Tao 2019). Our model consistently outperforms all the approaches on most metrics, which highlights the importance of visual understanding from visual and semantic modules in visual dialogue. CoAtt and HeiCoAtt-QI are relevant to our model in the sense that they leverage attention mechanism to identify question-relevant visual features. However, they ignore the

Table 2: Comparison on test-standard split of VisDial v1.0.

Model	MRR	R@1	R@5	R@10	Mean	NDCG
LF	55.42	40.95	72.45	82.83	5.95	45.31
HRE	54.16	39.93	70.47	81.50	6.41	45.46
MN	55.49	40.98	72.30	83.30	5.92	47.50
LF-Att	57.07	42.08	74.82	85.05	5.41	40.76
MN-Att	56.90	42.43	74.00	84.35	5.59	49.58
CorefMN	61.50	47.55	78.10	88.80	4.40	54.70
VGNN	61.37	47.33	77.98	87.83	4.57	52.82
RvA	63.03	49.03	80.40	89.83	4.18	55.59
DL-61	62.20	47.90	<b>80.43</b>	<b>89.95</b>	4.17	<b>57.32</b>
<b>DualVD</b>	<b>63.23</b>	<b>49.25</b>	80.23	89.70	<b>4.11</b>	56.32

semantic-rich relationships and language priors. It should be noted that our model and the compared approaches all belong to single-step models. With the success of multi-step reasoning, ReDAN (Gan et al. 2019) achieves 1% boost over our model on most metrics. We believe that stacking our visual encoder to achieve multi-step visual understanding is a promising future work. DL-61 (Guo, Xu, and Tao 2019) is a two-stage network for candidate selection and re-ranking while FGA (Schwartz et al. 2019) conducts attention across all the data parts, which gain relatively high performance on some metrics compared with our model. We believe that our model for the visual part and existing works for the dialogue or answer parts have complementary advantages.

## Ablation Study

Ablation study on VisDial v1.0 validation set exploits the influence of the essential components of DualVD. We use the same discriminative decoder for all the following variations:

**Object Representation (ObjRep):** this model uses the averaged object features to represent an image. Object representations are enhanced by question-driven attention.

**Relation Representation (RelRep):** this model applies averaged relation-aware object representations via *question-guided relation attention* and *question-guided graph convolution* as the image representation.

**Visual Module without Relationships (VisNoRel):** this is our full visual module except that the relation embeddings are replaced by unlabeled edges and the convolution is conducted via the intra-modal attention (Gao et al. 2019).

**Visual Module (VisMod):** this is our full visual module, which fuses objects and relation features.

**Global Caption (GICap):** this model uses LSTM to encode the global caption to represent the image.

**Local Caption (LoCap):** this model uses LSTM to encode the local captions to represent the image.

**Semantic Module (SemMod):** this is our full semantic module, which fuses global and local features.

**DualVD (full model):** this is our full model, which incorporates both the visual module and semantic module.

In Table 3, models in the first block are designed to evaluate the influence of key components in the visual module. **ObjRep** only considers isolated objects and ignores the relational information, which achieves worse performance compared with VisMod. **RelRep** considers the relationships by introducing relation embedding. However, empirical study

Table 3: Ablation study of DualVD on VisDial v1.0.

Model	MRR	R@1	R@5	R@10	Mean	NDCG
ObjRep	63.84	49.83	81.27	90.29	4.07	55.48
RelRep	63.63	49.25	81.01	90.34	4.07	55.12
VisNoRel	63.97	49.87	81.74	90.60	4.00	56.73
VisMod	64.11	50.04	81.78	90.52	3.99	56.67
GICap	60.02	45.34	77.66	87.27	4.78	50.04
LoCap	60.95	46.43	78.45	88.17	4.62	51.72
SemMod	61.07	46.69	78.56	88.09	4.59	51.10
w/o ELMo	63.67	49.89	80.44	89.84	4.14	56.41
<b>DualVD</b>	<b>64.64</b>	<b>50.74</b>	<b>82.10</b>	<b>91.00</b>	<b>3.91</b>	<b>57.30</b>

indicates that enhancing visual relationships while weakening object appearance is still not sufficient for better performance. **VisNoRel** fuses the information from both object appearance and neighborhoods without relational semantics, which achieves slight improvement compared to ObjRep. On top of VisNoRel, **VisMod** moves a step further by aggregating all the neighborhood features with relational information, which achieves the best performance compared to above three models.

Orthogonal to visual part, models in the second block evaluate the influence of key components in the semantic part. The overall performance of either **GICap** or **LoCap** decreases by 1% and 0.15% respectively, compared to their integrated version **SemMod**, which adaptively selects and fuses the task-specific descriptive clues from both global-level and local-level captions.

**DualVD** results in a great boost compared to SemMod and a relatively slight boost compared to VisMod. This unbalanced boost indicates that visual module provides comparatively richer clues than semantic module. Combining the two modules together gains an extra boost because of their complementary information. The performance of DualVD without ELMo embedding decrease slightly, which proves that the improvement of DualVD mainly comes from the contribution of the novel visual representation.

## Interpretability

A critical advantage of DualVD lies in its interpretability: DualVD is capable to predict the attention weights in the visual module, semantic module and the gate values in visual-semantic fusion. It supports explicit visualization and can reveal DualVD’s mode in information selection. Figure 3 shows three examples with variant dependence on visual and semantic modules. The third example (third and fourth rows in Figure 3) shows three round of dialogues about an image. In each round of dialogue, DualVD is capable to capture the most relevant visual and semantic information regarding the current question. In the first question, the visual module highlights the face of a boy and the relationships to his body and the other boy, while the semantic module puts more attention on the captions describing the two boys, which all provide useful clues to infer the correct answer. In the second and third round of dialogues, DualVD respectively attends to the whole grass and the discs. In this example, the attended information is adaptively changed through the dialogue and this explains why the correct answer is selected.

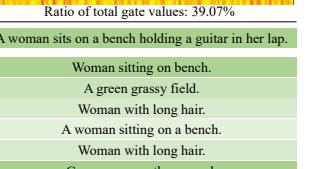
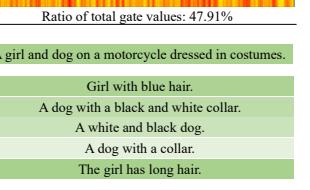
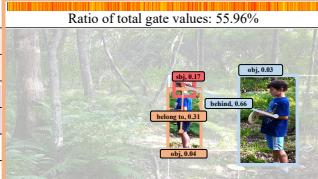
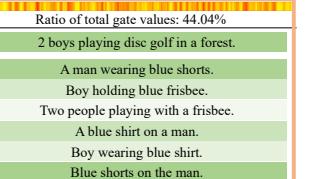
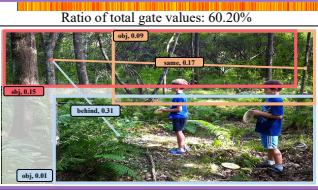
Image	Dialogue History	Visual Module	Semantic Module
	C: A woman sits on a bench holding a guitar in her lap. Q1: Is this in a park? A1: Yes, I believe it is. Q2: Are there others around? A2: No, she is alone. Q3: Does she have a collection bucket? A3: No. Q4: Is her hair long? A4: Yes, pretty long. Q5: Is she wearing a dress? A5: I don't think so, hard to tell. Q6: Does she have shoes on? A6: Yes, flip flops.	 Ratio of total gate values: 60.93%	 Ratio of total gate values: 39.07% A woman sits on a bench holding a guitar in her lap. Woman sitting on bench. A green grassy field. Woman with long hair. A woman sitting on a bench. Woman with long hair. Green grass on the ground.
	C: A girl and dog on a motorcycle dressed in costumes. Q1: Color pic? A1: Yes. Q2: Is the dog driving? A2: The dog is sitting on the motorcycle, like he would be the driver. Q3: Is the girl cute? A3: Yes. Q4: Is she old? A4: Around 9. Q5: Is she smart? A5: I'm not sure. Q6: Does she have hair? A6: Yes.	 Ratio of total gate values: 52.09%	 Ratio of total gate values: 47.91% A girl and dog on a motorcycle dressed in costumes. Girl with blue hair. A dog with a black and white collar. A white and black dog. A dog with a collar. The girl has long hair.
	C: 2 boys playing disc golf in a forest. Question1: Are the boys teenagers? Answer1: They are young boys. Question2: Do you see a lot of trees? Answer2: Yes, a ton of trees. Question3: Does one of the boys holding the disc? Answer3: They are both holding discs.	 Ratio of total gate values: 55.96%	 Ratio of total gate values: 44.04% 2 boys playing disc golf in a forest. A man wearing blue shorts. Boy holding blue frisbee. Two people playing with a frisbee. A blue shirt on a man. Boy wearing blue shirt. Blue shorts on the man.
	Ratio of total gate values: 60.20% Ratio of total gate values: 39.80%	 Ratio of total gate values: 54.90%	 Ratio of total gate values: 45.10% 2 boys playing disc golf in a forest. A man wearing blue shorts. Boy holding blue frisbee. Two people playing with a frisbee. A blue shirt on a man. Boy wearing blue shirt. Blue shorts on the man.

Figure 3: Visualization for DualVD. Visual module highlights the most relevant subject (red box) according to attention weights of each object ( $\gamma_i^v$  in Eq. 9) and the objects (orange and blue boxes) with the top two attended relationships ( $\beta_{ij}$  in Eq. 5). Semantic module shows the attention distribution ( $\delta_i^q$  in Eq. 11) over the global caption (first row) and the local captions (rest rows), where darker green color indicates bigger attention weight. The yellow thermogram on the top visualizes the gate values ( $gate^s$  in Eq. 16) of the visual embedding (left) and the caption embedding (right) in visual-semantic fusion. The ratio of gate values for the visual module and semantic module is computed from Eq. 16.

We further show another two examples with a current question and the dialogue history (first two rows in Figure 3) to reveal DualVD’s mode in information selection. We observe that the amount of information derived from each module highly depends on the complexity of the question and the relevance of the content. More information will come from the semantic module when the question involves complex relationships or the semantic module explicitly contains question-relevant clues. In Figure 3, *ratio of total gate values* reveals the amount of information derived from each module. In the first example, more visual information is required. Similar observation exists for the second question in the third example. Such questions referring to object appearance depend more clues from the visual module. In the second example, the current question is about the relationship between the girl and the hair. The amount of semantic information remarkably increases since there exists explicit evidence “*The girl has long hair*”. This observation holds for the third question in the third example. Since language is a higher-level encoding of the visual content after complex reasoning involved with prior knowledge, it provides more useful clues for semantic-level questions.

## Conclusion

In this paper, inspired by the dual-coding theory in cognitive science, we propose a novel DualVD model for visual dialogue. DualVD mainly consists of a visual module and a semantic module, which encodes image information at appearance-level and semantic-level, respectively. Desired clues for answer inference are adaptively selected from the two modules via gate mechanism. Results from extensive experiments on benchmarks demonstrate that deriving visual information from visual-semantic representations can achieve superior performance compared to other state-of-the-art approaches. Another major advantage of DualVD is its interpretability via progressive visualization. It can give us insight of how information from different modalities is used for inferring answers.

## Acknowledgement

This work is supported by the National Key Research and Development Program (Grant No.2017YFB0803301).

## References

- [Agrawal et al. 2017] Agrawal, A.; Lu, J.; Antol, S.; Mitchell, M.; Zitnick, C. L.; Parikh, D.; and Batra, D. 2017. Vqa: Visual question answering. *IJCV* 123(1):4–31.
- [Anderson et al. 2018] Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 6077–6086.
- [Das et al. 2017] Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M. F.; Parikh, D.; and Batra, D. 2017. Visual dialog. In *CVPR*, 1080–1089.
- [Gan et al. 2019] Gan, Z.; Cheng, Y.; Kholy, A. E.; Li, L.; and Gao, J. 2019. Multi-step reasoning via recurrent dual attention for visual dialog. In *ACL*.
- [Gao et al. 2018] Gao, P.; Li, H.; Li, S.; Lu, P.; Li, Y.; Hoi, S. C.; and Wang, X. 2018. Question-guided hybrid convolution for visual question answering. In *ECCV*, 469–485.
- [Gao et al. 2019] Gao, P.; Jiang, Z.; You, H.; Lu, P.; Hoi, S. C.; Wang, X.; and Li, H. 2019. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *CVPR*, 6639–6648.
- [Guo, Xu, and Tao 2019] Guo, D.; Xu, C.; and Tao, D. 2019. Image-question-answer synergistic network for visual dialog. In *CVPR*, 10434–10443.
- [Hudson and Manning 2019] Hudson, D. A., and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 6700–6709.
- [Jain, Lazebnik, and Schwing 2018] Jain, U.; Lazebnik, S.; and Schwing, A. 2018. Two can play this game: Visual dialog with discriminative question generation and answering. In *CVPR*, 5754–5763.
- [Johnson, Karpathy, and Fei-Fei 2016] Johnson, J.; Karpathy, A.; and Fei-Fei, L. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, 4565–4574.
- [Kottur et al. 2018] Kottur, S.; Moura, J. M.; Parikh, D.; Batra, D.; and Rohrbach, M. 2018. Visual coreference resolution in visual dialog using neural module networks. In *ECCV*, 153–169.
- [Li et al. 2019a] Li, H.; Wang, P.; Shen, C.; and Hengel, A. V. D. 2019a. Visual question answering as reading comprehension. In *CVPR*, 6319–6328.
- [Li et al. 2019b] Li, L.; Gan, Z.; Cheng, Y.; and Liu, J. 2019b. Relation-aware graph attention network for visual question answering. In *ICCV*.
- [Lu et al. 2016] Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical question-image co-attention for visual question answering. In *NIPS*, 289–297.
- [Lu et al. 2017] Lu, J.; Kannan, A.; Yang, J.; Parikh, D.; and Batra, D. 2017. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. In *NIPS*, 314–324.
- [Niu et al. 2019] Niu, Y.; Zhang, H.; Zhang, M.; Zhang, J.; Lu, Z.; and Wen, J.-R. 2019. Recursive visual attention in visual dialog. In *CVPR*, 6679–6688.
- [Paivio 1971] Paivio, A. 1971. *Imagery and Verbal Process*. New York: Holt, Rinehart and Winston.
- [Pennington, Socher, and Manning 2014] Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *EMNLP*, 1532–1543.
- [Peters et al. 2018] Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *NAACL*.
- [Qi et al. 2017] Qi, W.; Chunhua, S.; Peng, W.; Anthony, D.; and Anton van den, H. 2017. Image captioning and visual question answering based on attributes and external knowledge. *TPAMI* 40(6):1367–1381.
- [Qi et al. 2018] Qi, W.; Peng, W.; Chunhua, S.; Reid, I.; and Anton van den, H. 2018. Are you talking to me? reasoned visual dialog generation through adversarial learning. In *CVPR*, 6106–6115.
- [Ren et al. 2017] Ren, S.; Girshick, R.; Girshick, R.; and Sun, J. 2017. Faster r-cnn: Towards real-time object detection with region proposal networks. *TPAMI* 39(6):1137–1149.
- [Ren, Kiros, and Zemel 2015] Ren, M.; Kiros, R.; and Zemel, R. 2015. Exploring models and data for image question answering. In *NIPS*, 2953–2961.
- [Schwartz et al. 2019] Schwartz, I.; Yu, S.; Hazan, T.; and Schwing, A. G. 2019. Factor graph attention. In *CVPR*, 2039–2048.
- [Seo et al. 2017] Seo, P. H.; Lehrmann, A.; Han, B.; and Sigal, L. 2017. Visual reference resolution using attention memory for visual dialog. In *NIPS*, 3719–3729.
- [Simonyan and Zisserman 2014] Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *Computer Science*.
- [Wang et al. 2019] Wang, P.; Wu, Q.; Cao, J.; Shen, C.; Gao, L.; and Hengel, A. v. d. 2019. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *CVPR*, 1960–1968.
- [Xu and Saenko 2016] Xu, H., and Saenko, K. 2016. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*, 451–466.
- [Yang et al. 2016] Yang, Z.; He, X.; Gao, J.; Li, D.; and Smola, A. 2016. Stacked attention networks for image question answering. In *CVPR*, 21–29.
- [Yang, Zha, and Zhang 2019] Yang, T.; Zha, Z.-J.; and Zhang, H. 2019. Making history matter: Gold-critic sequence training for visual dialog. *arXiv preprint arXiv:1902.09326*.
- [Zhang et al. 2019] Zhang, J.; Kalantidis, Y.; Rohrbach, M.; Paluri, M.; Elgammal, A.; and Elhoseiny, M. 2019. Large-scale visual relationship understanding. In *AAAI*.
- [Zheng et al. 2019] Zheng, Z.; Wang, W.; Qi, S.; and Zhu, S.-C. 2019. Reasoning visual dialogs with structural and partial observations. In *CVPR*, 6669–6678.