$\mathbb{VD}$-$\mathbb{GR}$: Boosting $\mathbb{V}$isual $\mathbb{D}$ialog with Cascaded Spatial-Temporal Multi-Modal $\mathbb{GR}$aphs

Adnen Abdessaied   Lei Shi   Andreas Bulling

Institute for Visualization and Interactive Systems, University of Stuttgart, Germany

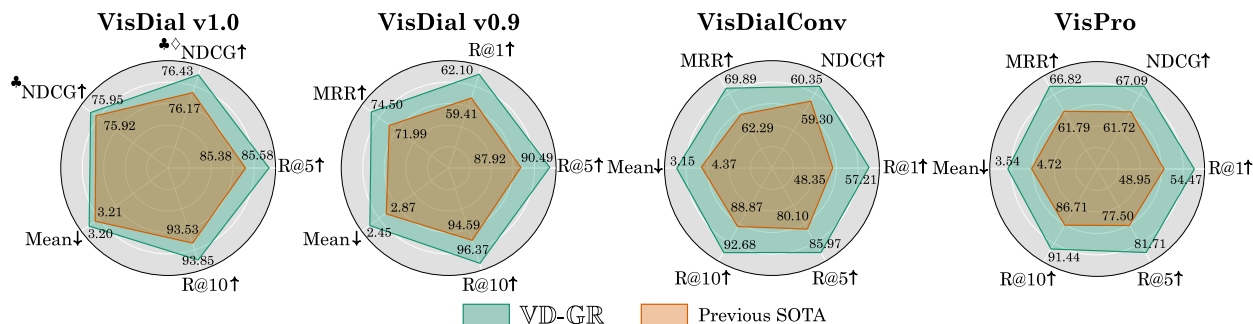{adnen.abdessaied, lei.shi, andreas.bulling}@vis.uni-stuttgart.de

Figure 1. $\mathbb{VD}$-$\mathbb{GR}$ outperforms strong baselines and achieves new state-of-the-art results on VisDial v1.0, VisDial v0.9, VisDialConv, and VisPro. ↑ indicates higher is better and ↓ indicates lower is better. (♣ = Fine-tuning on dense annotations, ◇ = Ensemble model).

Abstract

We propose $\mathbb{VD}$-$\mathbb{GR}$ – a novel visual dialog model that combines pre-trained language models (LMs) with graph neural networks (GNNs). Prior works mainly focused on one class of models at the expense of the other, thus missing out on the opportunity of combining their respective benefits. At the core of $\mathbb{VD}$-$\mathbb{GR}$ is a novel integration mechanism that alternates between spatial-temporal multi-modal GNNs and BERT layers, and that covers three distinct contributions: First, we use multi-modal GNNs to process the features of each modality (image, question, and dialog history) and exploit their local structures before performing BERT global attention. Second, we propose hub-nodes that link to all other nodes within one modality graph, allowing the model to propagate information from one GNN (modality) to the other in a cascaded manner. Third, we augment the BERT hidden states with fine-grained multi-modal GNN features before passing them to the next $\mathbb{VD}$-$\mathbb{GR}$ layer. Evaluations on VisDial v1.0, VisDial v0.9, VisDialConv, and VisPro show that $\mathbb{VD}$-$\mathbb{GR}$ achieves new state-of-the-art results across all four datasets.

🌐 Our project web-page is accessible here.

1. Introduction

Visual dialog is a multi-modal task to assess how well an artificial agent can hold a conversation with a human on a visual content using natural language [8]. Visual dialog differs from other tasks, such as visual [3] or video question answering [49], in that it requires the agent to answer a series of *temporally dependent* questions. That is, the agent not only has to reason about the visual input but also has to leverage the context of previous rounds to be able to answer the current question correctly. Although other datasets have been proposed for this task [22, 39], VisDial [8] has established itself as the de-facto standard because of its challenging, open-ended, and real-world nature.

Early visual dialog models on this dataset were based on recurrent networks [15] within deep neural architectures ranging from vanilla LSTMs [14] over memory nets [48] to hierarchical structures [40]. More recently, graph neural networks (GNNs) have been proposed and have been shown to produce more fine-grained features based on the local structures of each modality [6, 12, 16, 17, 25]. Other works have focused on attention models and pre-trained language models (LM) [1, 8, 33, 38], fine-tuned for the visual dialog task [27, 32, 46, 55]. However, both methodological approaches have so far remained separate despite their com-

plementary strengths and weaknesses: While GNNs are effective at exploiting local structure, they struggle to capture the global inter-modal context, especially in a rich multi-modal task like visual dialog. In contrast, transformer-based models [44] (e.g. BERT [9]) excel at learning the global context using self and global attention but often fail to exploit local intra-modal structures within each modality and suffer from the lack of inductive bias [50].

We posit that it is essential to exploit both local intra-modal structures and the global inter-modal context: Each modality is composed of smaller entities whose relationships have to be separately captured and understood by the model (i.e. the objects in the image, the words that constitute the question, and the rounds that form the history). At the same time, the inter-dependency of the modalities and the global context are equally important for a model to be able to answer the current question efficiently. Implementing this idea, we propose $\mathbb{VD}$-$\mathbb{GR}$ – the first visual dialog model to combine transformer-based LMs and GNNs. Each of our proposed layers alternates between GNNs that use hub-nodes to propagate information from one modality graph to another to alleviate the lack of inter-modal context, as well as BERT layers to learn the global intra-model context. The contributions of our work are threefold:

• A novel integration method of GNNs into transformer-based models that alternates between multi-modal graph aggregation and BERT layers. The GNNs exploit the local structure of each modality to augment the BERT hidden states with their fine-grained features before passing them to the next $\mathbb{VD}$-$\mathbb{GR}$ layer in an attempt to mitigate their lack of inductive bias.

• A novel feature propagation technique for multi-modal GNNs that relies on hub-nodes that link to all other nodes of the other modality in a cascaded manner, thus alleviating the lack of inter-modal context within the multi-modal graphs.

• Our model achieves new state-of-the-art results on VisDial v1.0, VisDial v0.9, VisDialConv, and VisPro datasets, thereby outperforming strong baselines on all four datasets.

2. Related Work

2.1. GNNs for Visual Dialog

GNN-EM [59] was one of the earliest models to deploy GNNs for visual dialog. It relied on building graphically structured neural networks to approximate the learning and inference processes of graphical models [10, 28, 29, 42]. FGA [38] developed a general attention mechanism that borrowed from the core idea of GNN message passing, and that was able to combine representations of any number of utilities. CAG [12] proposed representing the image as a fully-connected graph of objects whose adjacency

matrices could be dynamically updated using a question-based attention mechanism. GOG [6] proposed representing each modality as a graph before processing them by a light-weight fusion module [33] to rank the candidate answers. Similarly to CAG [12], DualVD [18] proposed representing the image as a graph consisting of connected objects. They argued that the visual view helped to capture the appearance-level information, including objects and their relationships, while the semantic view enabled the agent to understand high-level visual semantics from the whole image to the local regions [18].

Our model differs from the aforementioned works in two major aspects: (1) Instead of concatenating the GNN node features, we propose using hub-nodes that link to all remaining nodes within a given modality. We argue that this improves the feature fusion on a local scale given that GNNs can freely learn how to integrate information extracted from previous modalities. (2) We marry GNNs with pre-trained LMs (BERT) in a novel way to benefit from the advantages of both worlds. Contrarily to [51,53] where simple integration approaches for plain uni-modal tasks were introduced, we propose a sophisticated integration method for the complex and rich multi-modal visual dialog task: We first exploit the local multi-modal structures through GNNs before propagating inter-modal information in a cascaded manner via hub-nodes to finally enhance the hidden states of each BERT layer. To the best of our knowledge, we are the first to propose this for the visual dialog task.

2.2. Language Models for Visual Dialog

ViLBERT [30] and XLMERT [43] were among the first attempts to leverage pre-trained LMs for vision-language tasks (e.g. VQA [3], VCR [58], and image retrieval [54]). Concretely, they used a transformer encoder as a backbone and deployed a two-stream architecture to separately encode text and visual input. VisDial-BERT [32] built on top of ViLBERT and specifically adapted it to the visual dialog task, achieving new state-of-the-art performance on VisDial, and thus becoming the standard baseline for this dataset. More recently, UTC [5] was introduced as an improvement of VisDial-BERT and used two inter-task contrastive losses to improve training. Another class of models used a single stream to encode the multi-modal input such as B2T2 [2], VisualBERT [26], VL-BERT [41], and UNITER [7]. VD-BERT [46] adapted a single stream VQA model [60] to the visual dialog task and managed to achieve good performance without relying on external datasets.

Contrarily to the aforementioned works, our model relies on the power of cascaded GNNs to exploit local structures within each modality and, thus produce more fine-grained representations for subsequent BERT layers. To the best of our knowledge, this combination of GNNs and pre-trained LMs has not been explored before for visual dialog.
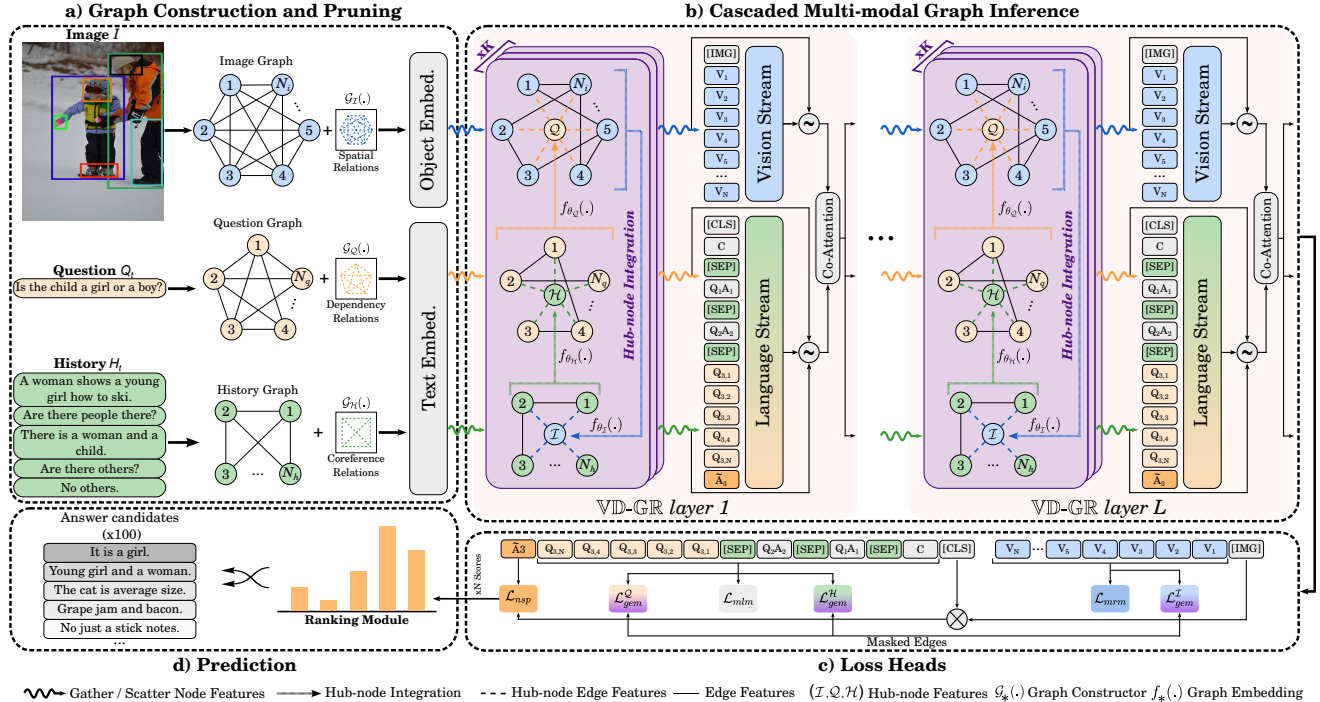
Figure 2. $\mathbb{VD}\text{-}\mathbb{GR}$ takes an image, a question, and a history consisting of the previous turns as input together with their respective graph structures. Each $\mathbb{VD}\text{-}\mathbb{GR}$-layer is composed of cascaded multi-modal graphs and a vanilla BERT layer.

3. Method

As shown in Figure 2, $\mathbb{VD}\text{-}\mathbb{GR}$ consists of four main components: a) a graph construction and pruning module, (b) a backbone operating on alternating cascaded multi-modal graphs and BERT layers, (c) multiple loss heads (next sentence prediction $\mathcal{L}_{nsp}$, masked language modelling $\mathcal{L}_{mlm}$, masked region modelling $\mathcal{L}_{mrm}$, and multi-modal graph edge masking ($\mathcal{L}_{gem}^{\mathcal{I}}$, $\mathcal{L}_{gem}^{\mathcal{Q}}$, $\mathcal{L}_{gem}^{\mathcal{H}}$)) , and (d) a prediction module to rank a set of candidate answers.

3.1. Problem Formulation

Given a question Q_t grounded on an image I at t-th turn, as well as its dialog history $H_t = \{C, (Q_1, A_1), ..., (Q_{t-1}, A_{t-1})\}$ (where C denotes the image caption), the model is tasked to predict its answer A_t by ranking a list of $N = 100$ answer candidates $\{\hat{A}_t^1, \hat{A}_t^2, ..., \hat{A}_t^{100}\}$.

3.2. Graph Construction and Pruning

Image Modality. The image graph constructor $\mathcal{G}_{\mathcal{I}}(.)$ treats each object in the image as a node and relies on spatial relationships to construct the graph topology, i.e. the adjacency matrix, where each edge represents a relationship between two objects in the image. The objects $\mathbf{I} = \{\mathbf{v}_1, ..., \mathbf{v}_{N_i}\}$ are obtained using Faster R-CNN [37] pre-trained on Visual Genome [23] where each object feature $\mathbf{v}_i$ is a 2048

dimensional vector and $N_i = 36$. Similar to [52], we distinguish between 11 relations based on the (x_1, y_1, x_2, y_2) object coordinates. Specifically, the overlapping region and spatial coordinates of two regions are used to judge whether an edge exists between them or not.

Question Modality. The question graph constructor $\mathcal{G}_{\mathcal{Q}}(.)$ treats each word in the question as a node and relies on dependency relations to construct the graph topology, i.e. the adjacency matrix, where each edge represents a relationship between two words in the question. To extract these relationships, we use the neural dependency parser of Stanza [36] that yields 47 relations.

History Modality. The history graph constructor $\mathcal{G}_{\mathcal{Q}}(.)$ treats each dialog round as a node and relies on coreference relations to construct the graph topology, i.e. the adjacency matrix, where each edge represents a relationship between two rounds in the history.

As can be seen in Figure 2, the topology of all multi-modal graphs is computed *once* during a pre-processing stage and is always kept constant over time. We refer to the supplementary material for additional details.

3.3. Proposed Layer

3.3.1 Transformer Features

Inline with previous works, we use VisDial-BERT [32] layers within our novel GNN-enhanced approach. For a given

image-question pair at round t, we first concatenate the caption C, the previous dialog rounds $\mathtt{H_t}$, the current question $\mathtt{Q_t}$, and a candidate answer $\mathtt{\tilde{A}_t}$ to form the textual input

$$\mathbf{T} = \left\{ \texttt{[CLS] C [SEP] Q\_1 [SEP]A\_1,..,Q\_t [SEP]} \, \tilde{A}_t \right\}, \tag{1}$$

where [CLS] and [SEP] are the special classification and separation tokens, respectively. As in [5,32,46,55], we use a special learnable token [IMG], and initialise it using mean pooling of the object sequence. The hidden features of the [IMG] token are used in conjunction with those of [CLS] by means of element-wise multiplication to produce the final input features of the NSP head.

Finally, the two sequences are used as input for the language and vision streams of the VisDial-BERT layer to obtain the hidden states $\mathbf{T}_h^{(l)}$ and $\mathbf{I}_h^{(l)}$, where l is the $\mathbb{VD}$-$\mathbb{GR}$ layer index.

3.3.2 Spatial-Temporal GNN Features

Node Features. As illustrated in Figure 2, the multi-modal GNNs of the l-th $\mathbb{VD}$-$\mathbb{GR}$ layer get their node features from the hidden states of the previous VisDial-BERT layer (or from the embedding layers in the first step). It is worth noting that the special tokens [IMG] and [CLS] are not included in the graph features. The nodes of the image and question graphs are gathered from the image and question token embeddings of the $(l-1)$-th layer, i.e. $\{\mathbf{I}_{h,i}^{(l-1)}\}_{i=1}^{N_i}$ and $\{\mathbf{T}_{h,i}^{(l-1)}\}_{i=s_q}^{s_q+N_q}$, where s_q and N_q denote the index of the first question token and the length of the question, respectively. We use the special [SEP] tokens to represent each dialog round in the history and gather their hidden states from $\mathbf{T}_h^{(l-1)}$ to get the node features of the history graph. The node features of the question and history graphs have to be extracted carefully since the textual input, i.e. where the question starts and ends and where the [SEP] tokens are located, varies within $\mathbf{T}_h^{(l-1)}$ for each dialog and round.

Hub-node Features. To make each modality aware of the other in a cascaded manner, we introduce hub-nodes as illustrated in Figure 2. As a result, the history becomes aware of the image, the question becomes aware of the history, and finally the image becomes aware of the question. The hub-node within each graph links to all other nodes using a special edge feature to propagate information on a *local scale* from one modality to the other before applying self- and cross-attention. To obtain the hub-node features, we train attention-based graph embeddings for each modality, i.e. $f_{\theta_\mathcal{I}}(.)$, $f_{\theta_\mathcal{Q}}(.)$, and $f_{\theta_\mathcal{H}}(.)$, which take the node features of the corresponding graph and output a single vector representing it. For example, the image hub-node $\mathcal{I}$, is computed from the image graph node features $\mathbf{I}_G$ of the l-th layer as

follows:

$$\mathcal{I} = f_{\theta_\mathcal{I}}(\mathbf{I}_G) = \sum_i \alpha_i \mathbf{v}_i \text{ for } \mathbf{v}_i \in \mathbf{I}_G, \tag{2}$$

$$\alpha = \{\alpha\}_i = \mathrm{MLP}(\mathbf{I}_G), \tag{3}$$

where MLP is a multi-layer perceptron that maps from the node features' vector space to $\mathbb{R}$ and $\theta_\mathcal{I}$ is the set of learnable parameters of the embedding. The question hub-node $\mathcal{Q}$ and history hub-node $\mathcal{H}$ are obtained in the same manner.

3.3.3 Graph Aggregation

Our multi-modal graphs are a variant of graph attention networks [45], although other types of GNNs can be used within each $\mathbb{VD}$-$\mathbb{GR}$ layer. Each multi-modal graph uses K layers (purple boxes in Figure 2) to propagate information between the nodes. For example, the k-th image graph layer receives the node features from the previous graph layer and updates them as follows:

$$\mathbf{I}_G^{(k)} = \{\mathbf{v}_1^{(k)}, .., \mathbf{v}_{N_i}^{(k)}, \underbrace{\mathbf{v}_{N_i+1}^{(k)}}_{=\mathcal{Q}}\} = \mathrm{GNN}_\mathcal{I}(\mathbf{I}_G^{(k-1)}), \tag{4}$$

$$\mathbf{v}_i^{(k)} = \mathrm{GeLU}\left(\mathop{\|}_{h=1}^{H} \tilde{\mathbf{v}}_i^{(k,h)} + \mathbf{v}_i^{(k-1)} \right), \tag{5}$$

$$\tilde{\mathbf{v}}_i^{(k,h)} = f\Big(\sum_{v_s \in \mathcal{N}_{v_t}} \alpha_{s \to t}^{(k,h)} \mathbf{m}_{s \to t}^{(k,h)} \Big) \, \forall 1 \le h \le H, \tag{6}$$

where $\|$ and H are the concatenation operation and the number of GNN attention heads, respectively. $\mathcal{N}_{v_t}$ represents the neighbourhood of node v_t, $\alpha_{s \to t}$ denotes the attention weight that scales the message $\mathbf{m}_{s \to t}^{(k,h)}$ between a source node $\mathbf{v}_s$ and a target node $\mathbf{v}_t$, and f is a linear layer. The messages $\mathbf{m}_{s \to t}^{(k,h)}$ between the nodes are computed following:

$$\mathbf{m}_{s \to t}^{(k,h)} = g_h(\mathbf{v}_s^{(k-1)}, \mathbf{e}_{s \to t}), \tag{7}$$

where g_h is a linear layer and $\mathbf{e}_{s \to t}$ is the edge feature between the nodes $\mathbf{v}_s$ and $\mathbf{v}_t$. We omitted the l index of the $\mathbb{VD}$-$\mathbb{GR}$ layer in the previous equations for brevity. The k-th question and history graph layers update their node features $\mathbf{Q}_G^{(k)}$ and $\mathbf{H}_G^{(k)}$ in the same manner.

3.3.4 Hidden States Enhancement

The outputs of the multi-modal graphs of the last K-th layer, i.e. $\mathbf{I}_G^{(l,K)}$, $\mathbf{Q}_G^{(l,K)}$ and $\mathbf{H}_G^{(l,K)}$, are used to enhance the hidden states of the following BERT layer. First, the GNN features are scattered back to their corresponding places within the VisDial-BERT hidden states. Then, we apply a fusion operation inspired by the idea of residual connections [13] as illustrated in Figure 2. Our experiments show

Method	VisPro						VisDialConv					
	NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓	NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
MCA-I [1]	59.80	57.88	45.39	72.24	82.76	5.84	52.07	55.55	41.65	72.47	83.81	5.92
MCA-I-HConcQ [1]	61.08	61.79	48.95	77.50	86.58	4.72	54.84	62.06	47.42	80.10	88.87	4.37
MCA-I-HGuidedQ [1]	61.35	60.13	47.11	75.26	86.18	5.23	53.81	62.29	48.35	80.10	88.76	4.42
MCA-I-VGH [1]	61.68	59.33	46.18	75.53	86.71	5.07	55.48	58.48	44.54	74.95	86.19	5.18
MCA-I-H [1]	61.72	59.62	45.92	77.11	86.45	4.85	53.01	61.24	47.63	79.07	87.94	4.77
Student [19]	–	–	–	–	–	–	59.30	–	–	–	–	–
VD-GR	67.09	66.82	54.47	81.71	91.44	3.54	60.35	69.89	57.21	85.97	92.68	3.15

Table 1. Performance comparison on VisPro. The best and second-best results are in **bold** and underlined, respectively. ↑ indicates higher is better and ↓ indicates lower is better.

that this step is crucial and leads to significant improvements in performance. These operations can be summarised as follows:

$$\tilde{\mathbf{I}}_h^l = \mathbf{I}_h^l \oslash (\mathbf{I}_G^{(l,K)}, \mathrm{Idx}_v), \;\; \hat{\mathbf{I}}_h^l = \lambda \mathbf{I}_h + (1-\lambda)\tilde{\mathbf{I}}_h^l, \quad (8)$$

$$\tilde{\mathbf{T}}_h^l = \left(\mathbf{T}_h^l \oslash (\mathbf{Q}_G^{(l,K)}, \mathrm{Idx}_q)\right) \oslash \left(\mathbf{H}_G^{(l,K)}, \mathrm{Idx}_h\right), \quad (9)$$

$$\hat{\mathbf{T}}_h^l = \lambda \mathbf{T}_h + (1-\lambda)\tilde{\mathbf{T}}_h^l \quad \text{for } \lambda \in [0,1], \quad (10)$$

where $\oslash$ denote the scatter operation and Idx_* the indices of the graph nodes features with respect to the BERT hidden states. The final enhanced features $\hat{\mathbf{I}}_h^l$ and $\hat{\mathbf{T}}_h^l$ are passed to the next BERT layer.

3.4. Loss Heads

We complement the traditional losses used for the visual dialog task (masked language modelling $\mathcal{L}_{\mathrm{mlm}}$, masked region modelling $\mathcal{L}_{\mathrm{mrm}}$, and next sentence prediction $\mathcal{L}_{\mathrm{nsp}}$) with graph edge masking $\mathcal{L}_{\mathrm{gem}}$ to improve learning of the local structure of each modality, and thus to enhance the feature representation of our multi-modal graphs.

Masked Language and Region Modelling. Similar to masked language modelling introduced in [9], we randomly masked 10% of the text tokens and image objects with the special token [MASK] and the model had to recover them based on the surrounding tokens and cross-modal clues:

$$\mathcal{L}_{\mathrm{mlm}} = -\mathbb{E}_{(\mathbf{w},\mathbf{I})\sim S_{tr}} \left[log P(w_m|\mathbf{w}_{\backslash m}, \mathbf{h}_{[\texttt{IMG}]}) \right], \quad (11)$$

$$\mathcal{L}_{\mathrm{mrm}} = -\mathbb{E}_{(\mathbf{w},\mathbf{I})\sim S_{tr}} \left[log P(\mathbf{v}_m|\mathbf{I}_{\backslash m}, \mathbf{h}_{[\texttt{CLS}]}) \right], \quad (12)$$

where w_m, $\mathbf{v}_m$, and S_{tr} denote the masked words, the masked regions, and the training set, respectively.

Next Sentence Prediction. The aim of next sentence prediction (NSP) is to identify whether the appended answer $\tilde{\mathtt{A}}_t$ is correct or not, i.e.

$$\mathcal{L}_{\mathrm{nsp}} = -\mathbb{E}_{(\mathbf{w},\mathbf{I})\sim S_{tr}} \left[log P(y|\aleph(\mathbf{w},\mathbf{I})) \right], \quad (13)$$

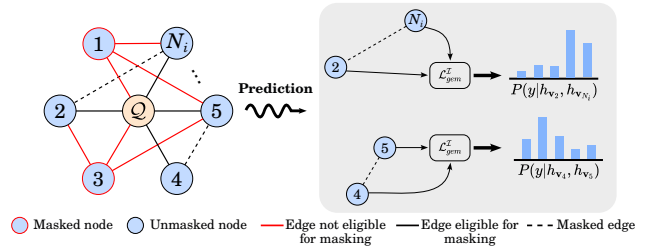

Figure 3. Only edges connecting two unmasked node features are eligible for masking with a 15% probability.

where $y \in \{0,1\}$ is a ground-truth binary label, and $\aleph(.,.)$ is the binary answer prediction head operating on the element-wise product of the [IMG] and [CLS] token representations.

Graph Edge Masking. We introduce the multi-modal graph edge masking to alleviate the potential problem of having well-trained BERT layers but not sufficiently trained GNNs. Given that, by design, our GNNs receive their features from the previous BERT layer, they inherit some masked node representations. In order to make the edge prediction task stable, we only masked 15% of the edges connecting two unmasked node features as illustrated in Figure 3. The representations of these nodes were then used to predict the masked edges. For the example of the image graph, this results in the following loss:

$$\mathcal{L}_{\mathrm{gem}}^{\mathcal{I}} = -\mathbb{E}_{(\mathbf{w},\mathbf{I})\sim S_{tr}} \left[log P(y_m^{(i,j)}|\mathbf{h}_{\mathbf{v}_i}, \mathbf{h}_{\mathbf{v}_j}) \right], \quad (14)$$

where $y_m^{(i,j)}$ is the ground-truth edge type between the nodes $\mathbf{v}_i$ and $\mathbf{v}_j$. The question and history edge graph masking losses $\mathcal{L}_{\mathrm{gem}}^{\mathcal{Q}}$ and $\mathcal{L}_{\mathrm{gem}}^{\mathcal{H}}$ are obtained in a similar manner.

Total Loss. We adopt a two-stage approach to train our model. First, we train it on a warm-up task of masked token and graph edge prediction, i.e. using the total loss $\mathcal{L}_{\mathrm{warm}}$:

Method	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
MN [8]	59.65	45.55	76.22	85.37	5.46
CoAtt [57]	63.98	50.29	80.71	88.81	4.47
HCIAE [31]	62.22	48.48	78.75	87.59	4.81
CorefNMN [21]	64.10	50.92	80.18	88.81	4.45
RvA [34]	66.34	52.71	82.97	90.73	3.93
Student [19]	60.03	50.40	70.74	77.15	12.13
DVAN [11]	66.67	53.62	82.85	90.72	3.93
VD-BERT [46]	70.04	57.79	85.34	92.68	4.04
VisDial-BERT [32]	71.99	59.41	87.92	94.59	2.87
$\mathbb{VD}$-$\mathbb{GR}$	**74.50**	**62.10**	**90.49**	**96.37**	**2.45**

Table 2. Performance comparison on the *val* split of VisDial v0.9 dataset. NDCG is not supported in this version of the dataset.

$$\mathcal{L}_{\text{warm}} = \alpha_1(\mathcal{L}_{\text{mlm}} + \mathcal{L}_{\text{mrm}}) + \alpha_2 \mathcal{L}_{\text{GEM}}, \quad (15)$$

$$\mathcal{L}_{\text{GEM}} = \mathcal{L}_{\text{GEM}}^{\mathcal{I}} + \mathcal{L}_{\text{GEM}}^{\mathcal{Q}} + \mathcal{L}_{\text{GEM}}^{\mathcal{H}}. \quad (16)$$

Then, we only train the model based on the visual dialog loss

$$\mathcal{L}_{\text{VD}} = \mathcal{L}_{\text{mlm}} + \mathcal{L}_{\text{mrm}} + \mathcal{L}_{\text{nsp}}. \quad (17)$$

4. Experiments

4.1. Datasets

We evaluated $\mathbb{VD}$-$\mathbb{GR}$ on the challenging VisDial v0.9 and VisDial v1.0 datasets. VisDial v0.9 has circa 83k training and 40k validation dialogs. The more recent v1.0 version consists of about 123k, 2k, and 8k images for training, validation, and testing, respectively. Each image comes with a caption and 10 question-answer pairs; each question turn is associated with 100 candidate answers. The validation data and part of the training data of VisDial v1.0 provide dense annotations for the candidate answers. Furthermore, we evaluated our model on two additional datasets, i.e. VisPro [56] and VisDialConv [1].

4.2. Quantitative Results

VisDialConv & VisPro. First, we evaluated our model on VisPro [1] and VisDialConv which were introduced to verify the role of dialog history in answering the current question Q_t. We compared $\mathbb{VD}$-$\mathbb{GR}$ to the baselines introduced in [1] as well as the most recent Student model [19]. As can be seen from Table 1, $\mathbb{VD}$-$\mathbb{GR}$ significantly outperformed all MCA variants across all metrics on both datasets. Specifically, it increased the performance of the baselines by over 5 absolute points on NDCG and MRR on VisPro. On VisDialConv, $\mathbb{VD}$-$\mathbb{GR}$ increased the top performance by over 4 absolute points on the same metrics. Table 1 also shows that our model managed to surpass the Student model by over 1

[1]Same subset as in [1].

Method	NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
LTMI [33]	62.72	62.32	48.94	78.65	87.88	4.86
VD-BERT [46]	63.22	67.44	54.02	83.96	92.33	3.53
VisDial-BERT [32]	60.96	67.17	53.42	84.41	92.62	3.41
MCA [1]	60.27	64.33	51.12	80.91	89.65	4.24
UniMM-UL [47]	62.86	53.49	42.70	65.03	74.58	10.65
UTC [5]	63.22	68.58	55.48	85.38	93.20	3.28
Student [19]	**65.47**	53.19	43.08	64.09	71.51	14.34
VD-PCR [55]	64.16	69.71	56.79	85.82	93.64	3.15
$\mathbb{VD}$-$\mathbb{GR}$	64.32	**69.91**	**57.01**	**86.14**	**93.74**	**3.13**

Table 3. Performance comparison on the *val* split of VisDial v1.0 dataset.

NDCG absolute point although it was trained on circa 13M additional images.

VisDial v0.9. Second, we compared $\mathbb{VD}$-$\mathbb{GR}$ with the state of the art on the *val* split of VisDial v0.9. As can be seen from Table 2, our model significantly outperformed all previous models and achieved new state-of-the-art results across all metrics. Specifically, it outperformed pre-training methods such as VisDial-BERT and VD-BERT by a large margin. Even more importantly, it managed to also surpass more recent models such as Student. Specifically, $\mathbb{VD}$-$\mathbb{GR}$ improved the MRR and R@1 scores by over 2.5 absolute points compared to the second best model.

VisDial v1.0 *val*. We then compared $\mathbb{VD}$-$\mathbb{GR}$ with the state of the art on the *val* split of VisDial v1.0. As can be seen from Table 3, our model outperformed all previous models across all five sparse metrics. Specifically, it outperformed pre-training methods, such as VisDial-BERT and VD-BERT by a significant margin. Even more importantly, it managed to also surpass more recent models such as UniMM-UL, VD-PCR, UTC, and Student. $\mathbb{VD}$-$\mathbb{GR}$ improved MRR, R@1, and R@5 by over 0.2 absolute points compared to the second best VD-PCR model. Furthermore, we compared their performance on individual dialog rounds using the sparse metrics (MRR, R@1, R@5, and R@10). As can be seen from Figure 4, $\mathbb{VD}$-$\mathbb{GR}$ managed to outperform VD-PCR on almost all rounds of the dataset.

VisDial v1.0 *test-std*. Finally, we compared our model with state-of-the-art published baselines on the *test-std* split of the VisDial v1.0 dataset.
• *State-of-the-art Results on Sparse Metrics:* As can be seen from the first section of Table 4, $\mathbb{VD}$-$\mathbb{GR}$ lifted the state-of-the-art R@5, R@10, and Mean scores from 85.38, 93.53, and 3.21 achieved by VD-PCR to 85.58, 93.85, and 3.20, respectively. On the remaining metrics, our models performed on par with the state of the art. Specifically, it reached respective scores of 68.65 and 55.33 on MRR and R@1, only third to UTC and VD-PCR.
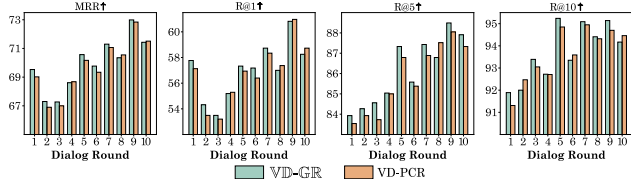
Figure 4. Performance comparison on different dialog rounds of the VisDial v1.0 *val* split. We only considered sparse metrics since the dense annotations used to compute the NDCG score are not defined on all 10 rounds of the validation dialogs.
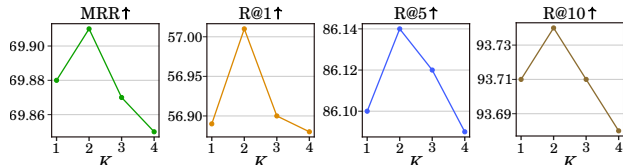

Figure 5. Performance comparison with different number of GNN layers K on the *val* split of VisDial v1.0. Results are shown for the *val* split of VisDial v1.0.

• *Fine-tuning on Dense Annotations:* As in previous works, we fine-tuned our model on the dense annotations released by [32] in order to improve the NDCG score. As illustrated in the second section of Table 4, the NDCG score sharply increased from 63.49 to 75.95, outperforming all previous models in the single-model setting. However, by fine-tuning on dense annotations, we decreased the performance on the sparse metrics (MRR, R@1, R@5, R@10, and Mean). This well-known phenomenon of the dataset is due to the misalignment of the dense and sparse annotations as previously observed by [32, 46]. In contrast to other baselines, $\mathbb{VD}$-$\mathbb{GR}$ managed to keep relatively high sparse scores (4/5 metrics are the new state of the art) after fine-tuning.

• *Ensemble Setting:* As it is common practice, we fine-tuned an ensemble comprised of eight pre-trained $\mathbb{VD}$-$\mathbb{GR}$ models to further improve the NDCG score. One set of four models was fine-tuned with Cross Entropy (CE) and with a varying number for GNN layers K, i.e. $K = \{1, 2, 3, 4\}$. Each model of the second set was trained using the List-Net [4] ranking optimisation method. As can be seen in the last section of Table 4, our ensemble model reached an NDCG score of 76.43 outperforming the closest competitor UniMM-UL with only 76.17 and, thus setting a new state of the art in the ensemble setting.

4.3. Ablation Study

Number of GNN layers K. This is an important hyperparameter of our model: If K is too small, then the expressive power of the GNNs will be hampered. Contrarily, if K is too large, the GNNs will suffer from over-smoothing [24]. To this end, we increased K incrementally from one to four

Method	NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
MN [8]	47.50	55.49	46.98	72.30	83.30	5.92
CorefNMN [21]	54.70	61.50	47.55	78.10	88.80	4.40
FGA [38]	56.90	66.20	52.75	82.92	91.07	3.80
DAN [20]	57.59	63.20	49.63	79.75	89.35	4.30
LTMI [33]	59.03	64.08	50.20	80.68	90.35	4.05
CAG [12]	56.64	63.49	49.85	80.63	90.15	4.11
GOG [6]	61.04	63.52	50.01	80.13	89.28	4.31
VD-BERT [46]	59.96	65.44	51.63	82.23	90.68	3.90
VisDial-BERT [32]	63.87	67.50	53.85	84.68	93.25	3.32
UTC [5]	64.60	68.70	**55.73**	84.93	93.08	3.32
VD-PCR [55]	63.55	**68.73**	55.45	85.38	93.53	**3.21**
UniMM-UL [47]	63.90	68.14	54.57	85.15	93.13	3.27
‡Student [19]	**64.91**	68.44	55.05	85.18	93.35	3.23
$\mathbb{VD}$-$\mathbb{GR}$	63.49	68.65	55.33	**85.58**	**93.85**	3.20
♣MCA [1]	72.47	37.68	20.67	56.67	72.12	8.89
♣VD-BERT [46]	74.54	50.74	33.15	61.58	77.15	7.18
♣VisDial-BERT [32]	74.47	50.74	37.95	64.13	80.00	6.28
♣UTC [5]	74.32	50.24	37.12	63.98	79.88	6.48
♣VD-PCR [55]	75.30	56.17	45.32	68.05	80.98	6.15
♣UniMM-UL [47]	75.92	56.18	43.70	71.03	**84.80**	5.42
♣$\mathbb{VD}$-$\mathbb{GR}$	**75.95**	**58.30**	**46.55**	71.45	84.52	**5.32**
◇♣P1+P2 [35]	74.91	49.13	36.68	62.96	78.55	7.03
◇♣VD-BERT [46]	75.35	51.17	38.90	62.82	77.98	6.69
◇♣VD-PCR [55]	76.14	56.05	44.75	68.40	82.75	5.72
◇♣UniMM-UL [47]	76.17	56.42	44.32	70.23	84.52	5.47
◇♣$\mathbb{VD}$-$\mathbb{GR}$	**76.43**	56.35	**45.18**	68.13	82.18	5.79

Table 4. Performance comparison on the *test-std* split of VisDial v1.0 dataset. ♣ indicates fine-tuning on dense annotations and ◇ denotes ensemble model. ‡ denotes the use of extra large datasets for training.

Ablation	NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
$\mathbb{VD}$-$\mathbb{GR}$ w/ $\lambda = 0$	56.69	67.71	54.07	85.03	92.84	3.33
$\mathbb{VD}$-$\mathbb{GR}$ w/o $\mathcal{L}_{\text{warm}}$	63.76	69.83	56.84	86.05	93.70	3.15
$\mathbb{VD}$-$\mathbb{GR}$ w/o sharing	64.15	69.79	56.73	86.02	93.68	3.15
$\mathbb{VD}$-$\mathbb{GR}$ w/o HN	64.11	69.86	56.88	85.97	93.67	3.14
$\mathbb{VD}$-$\mathbb{GR}$ (Full)	**64.32**	**69.91**	**57.01**	**86.14**	**93.74**	**3.13**

Table 5. Performance comparison of ablated versions of our model on the *val* split of VisDial v1.0.

and used the sparse metrics of the task (MRR, R@1, R@5, R@10) for validation. As illustrated in Figure 5, the performance of our model peaked at $K = 2$ for all four metrics. Thus, we kept this value fixed in all previous experiments unless explicitly stated otherwise.

Model Ablations. In addition to the full model, we evaluated the following ablated versions:

• $\mathbb{VD}$-$\mathbb{GR}$ **w/ $\lambda = 0$**: This variant did not apply the proposed residual connection of Equation 10 while augmenting the BERT hidden states.

• $\mathbb{VD}$-$\mathbb{GR}$ **w/o $\mathcal{L}_{\text{warm}}$**: This variant was not trained on the warm-up task of edge prediction as discussed in Sec. 3.4.

• $\mathbb{VD}$-$\mathbb{GR}$ **w/o sharing**: This variant did not share the weights of the GNNs in each layer of our model.

• $\mathbb{VD}$-$\mathbb{GR}$ **w/o HN**: This variant did not use hub-nodes to
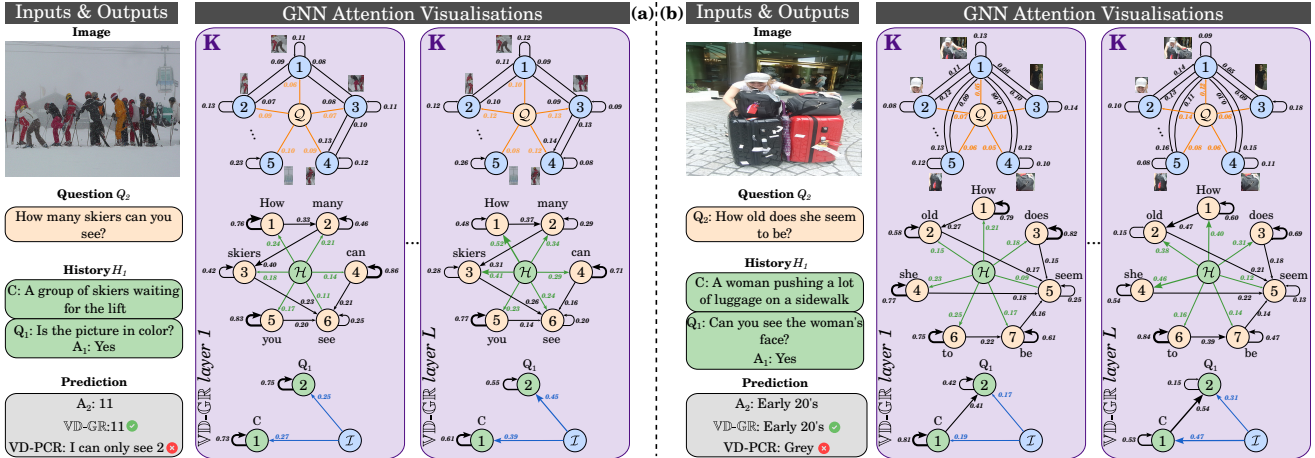
Figure 6. Qualitative samples from the VisDial v1.0 *val* split. The attention weights correspond to the K-th GNN of the first and last $\mathbb{VD}$-$\mathbb{GR}$ layers. We first average them across all H attention heads before re-normalising them for each node using a `softmax` function. For clarity, we only show a subset of the image graph nodes.

propagate the information between the multi-modal GNNs.

As can be seen from Table 5, the residual connections are essential for high performance. Without them, $\mathbb{VD}$-$\mathbb{GR}$ achieved the lowest performance across all metrics. The same applies to the warm-up training stage: Although the performance of this ablated version improved over the previous one, it still performed significantly worse than our best model. The results also underline the importance of sharing the GNN weights within the different $\mathbb{VD}$-$\mathbb{GR}$ layers: Although this version has more weights, it still performed worse than our best model on all metrics. This finding was to be expected given that the local structure of each modality does not change from one $\mathbb{VD}$-$\mathbb{GR}$ layer to another. Finally, the importance of the inter-modal feature propagation using hub-nodes is highlighted by the two last rows of Table 5: The hub-nodes enabled our model to achieve the best performance across all metrics of the *val* split of the VisDial v1.0 dataset.

4.4. Qualitative Results

Finally, in Figure 6 we show selected qualitative samples (with more in the supplementary material) from the *val* split of VisDial v1.0 alongside the ground truth answers, as well as the top-1 predictions of our $\mathbb{VD}$-$\mathbb{GR}$ model and VD-PCR for comparison since it achieved the second best results on this split. We make two interesting observations: (1) Our model deals better with questions that require exploiting local structure within modalities. For example, it managed to correctly answer $Q_2 =$ `How many skiers can you see?` in the first dialog sample (see Figure 6a) by predicting `11` whereas VD-PCR predicted `I can only see two`. We hypothesise that this is due to the fact that our model exploits the spatial structure of the visual input more ef-

fectively using the image graph compared to VD-PCR, although the latter has access to the same visual features. (2) $\mathbb{VD}$-$\mathbb{GR}$ has more accurate semantic understanding of the question. This is highlighted in answering $Q_2 =$ `How old does she seem to be?` (referring to the woman) in the second dialog sample (see Figure 6b). Whereas our model correctly predicted `Early 20's`, VD-PCR failed by answering `Grey`, which is not a semantically-sound response. We posit that this advantage of our model is related to the fine-grained features of the question graph.

5. Conclusion

In this work we proposed $\mathbb{VD}$-$\mathbb{GR}$ – a novel visual dialog model that combines pre-trained language models and GNNs. Specifically, $\mathbb{VD}$-$\mathbb{GR}$ alternates between multi-modal graphs and BERT layers, and augments the hidden states of the latter with the fine-grained features obtained by the former. $\mathbb{VD}$-$\mathbb{GR}$ propagates information from one modality graph to another in a cascaded manner using hub-nodes that link to all other nodes within each modality, thereby effectively alleviating the lack of inter-modal context. Extensive analyses underlined its effectiveness, while experiments on four challenging visual dialog datasets (VisDial v1.0, VisDial v0.9, VisDialConv, and VisPro) demonstrated its superior performance over existing methods.

References

[1] Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. History for Visual Dialog: Do we really need it? In *ACL*, 2020. 1, 5, 6, 7

[2] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. Fusion of Detected Objects in Text for Visual Question Answering. In *EMNLP*, 2019. 2

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *ICCV*, 2015. 1, 2

[4] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to Rank: From Pairwise Approach to Listwise Approach. In *ICML*, 2007. 7

[5] Cheng Chen, Yudong Zhu, Zhenshan Tan, Qingrong Cheng, Xin Jiang, Qun Liu, and Xiaodong Gu. UTC: A Unified Transformer with Inter-Task Contrastive Learning for Visual Dialog. In *CVPR*, 2022. 2, 4, 6, 7

[6] Feilong Chen, Xiuyi Chen, Fandong Meng, Peng Li, and Jie Zhou. GoG: Relation-aware graph-over-graph network for visual dialog. In *Findings of ACL*, 2021. 1, 2, 7

[7] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 2

[8] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. In *CVPR*, 2017. 1, 6, 7

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 2, 5

[10] Haoshu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. Learning Pose Grammar to Encode Human Body Configuration for 3D Pose Estimation. In *AAAI*, 2017. 2

[11] Dan Guo, Hui Wang, and Meng Wang. Dual visual attention network for visual dialog. In *IJCAI*, 2019. 6

[12] Dan Guo, Hui Wang, Hanwang Zhang, Zheng Jun Zha, and Meng Wang. Iterative Context-Aware Graph Inference for Visual Dialog. In *CVPR*, 2020. 1, 2, 7

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997. 1

[15] J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 1982. 1

[16] Tianling Jiang, Yi Ji, Chunping Liu, and Hailin Shao. Visual-textual alignment for graph inference in visual dialog. In *COLING*, 2020. 1

[17] Xiaoze Jiang, Siyi Du, Zengchang Qin, Yajing Sun, and Jing Yu. KBGN: Knowledge-Bridge Graph Network for Adaptive Vision-Text Reasoning in Visual Dialogue. In *ACM MM*, 2020. 1

[18] Xiaoze Jiang, Jing Yu, Zengchang Qin, Yingying Zhuang, Xingxing Zhang, Yue Hu, and Qi Wu. DualVD: An Adaptive Dual Encoding Model for Deep Visual Understanding in Visual Dialogue. In *AAAI*, 2020. 2

[19] Gi-Cheon Kang, Sungdong Kim, Jin-Hwa Kim, Donghyun Kwak, and Byoung-Tak Zhang. The Dialog Must Go On: Improving Visual Dialog via Generative Self-Training. In *CVPR*, 2023. 5, 6, 7

[20] Gi-Cheon Kang, Jaeseo Lim, and Byoung-Tak Zhang. Dual attention networks for visual reference resolution in visual dialog. In *EMNLP*, 2019. 7

[21] Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Visual coreference resolution in visual dialog using neural module networks. In *ECCV*, 2018. 6, 7

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

[23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 3

[24] Guohao Li, Matthias Muller, Ali Thabet, and Bernard Ghanem. DeepGCNs: Can GCNs Go As Deep As CNNs? In *ICCV*, 2019. 7

[25] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware Graph Attention Network for Visual Question Answering. In *ICCV*, 2019. 1

[26] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A Simple and Performant Baseline for Vision and Language. In *arXiv:1908.03557*, 2019. 2

[27] Mingxiao Li and Marie-Francine Moens. Modeling coreference relations in visual dialog. In *EACL*, 2021. 1

[28] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated Graph Sequence Neural Networks. In *ICLR*, 2015. 2

[29] Guosheng Lin, Chunhua Shen, Ian Reid, and Anton van den Hengel. Deeply Learning the Messages in Message Passing Inference. In *NeurIPS*, 2015. 2

[30] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViL-BERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*, 2019. 2

[31] Jiasen Lu, Anitha Kannan, , Jianwei Yang, Devi Parikh, and Dhruv Batra. Best of Both Worlds: Transferring Knowledge from Discriminative Learning to a Generative Visual Dialog Model. In *NeurIPS*, 2017. 6

[32] Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In *ECCV*, 2020. 1, 2, 3, 4, 6, 7

[33] Van Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. Efficient Attention Mechanism for Visual Dialog that Can Handle All the Interactions Between Multiple Inputs. In *ECCV*, 2020. 1, 2, 6, 7

[34] Yulei Niu, Hanwang Zhang, Manli Zhang, Jianhong Zhang, Zhiwu Lu, and Ji-Rong Wen. Recursive Visual Attention in Visual Dialog. In *CVPR*, 2019. 6

[35] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. Two causal principles for improving visual dialog. In *CVPR*, 2020. 7

[36] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In *ACL*, 2020. 3

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NeurIPS*, 2015. 3

[38] Idan Schwartz, Seunghak Yu, Tamir Hazan, and Alexander G. Schwing. Factor Graph Attention. In *CVPR*, 2019. 1, 2, 7

[39] Paul Hongsuck Seo, Andreas Lehrmann, Bohyung Han, and Leonid Sigal. Visual Reference Resolution using Attention Memory for Visual Dialog. In *NeurIPS*, 2017. 1

[40] Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. In *AAAI*, 2017. 1

[41] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *ICLR*, 2020. 2

[42] Sainbayar Sukhbaatar, Rob Fergus, et al. Learning multi-agent communication with backpropagation. In *NeurIPS*, 2016. 2

[43] Hao Tan and Mohit Bansal. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP*, 2019. 2

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *NeurIPS*, 2017. 2

[45] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. In *ICLR*, 2018. 4

[46] Yue Wang, Shafiq Joty, Michael R. Lyu, Irwin King, Caiming Xiong, and Steven C.H. Hoi. VD-BERT: A unified vision and dialog transformer with BERT. In *EMNLP*, 2020. 1, 2, 4, 6, 7

[47] Zihao Wang, Junli Wang, and Changjun Jiang. Unified multimodal model with unlikelihood training for visual dialog. In *ACM MM*, 2022. 6, 7

[48] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory Networks. In *NeurIPS*, 2014. 1

[49] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *CVPR*, 2016. 1

[50] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. In *NeurIPS*, 2021. 2

[51] Junhan Yang, Zheng Liu, Shitao Xiao, Chaozhuo Li, Defu Lian, Sanjay Agrawal, Amit Singh, Guangzhong Sun, and Xing Xie. GraphFormers: GNN-nested Transformers for Representation Learning on Textual Graph. In *NeurIPS*, 2021. 2

[52] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, 2018. 3

[53] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? In *NeurIPS*, 2021. 2

[54] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2014. 2

[55] Xintong Yu, Hongming Zhang, Ruixin Hong, Yangqiu Song, and Changshui Zhang. VD-PCR: Improving visual dialog with pronoun coreference resolution. *Pattern Recognition*, 2022. 1, 4, 6, 7

[56] Xintong Yu, Hongming Zhang, Yangqiu Song, Yan Song, and Changshui Zhang. What You See is What You Get: Visual Pronoun Coreference Resolution in Dialogues. In *EMNLP-IJCNLP*, 2019. 6

[57] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multimodal factorized bilinear pooling with co-attention learning for visual question answering. In *ICCV*, 2017. 6

[58] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From Recognition to Cognition: Visual Commonsense Reasoning. In *CVPR*, June 2019. 2

[59] Zilong Zheng, Wenguan Wang, Siyuan Qi, and Song-Chun Zhu. Reasoning Visual Dialogs with Structural and Partial Observations. In *CVPR*, 2019. 2

[60] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified Vision-Language Pre-Training for Image Captioning and VQA. In *AAAI*, 2020. 2