Multimodal Graph Transformer for Multimodal Question Answering

Xuehai He
UC Santa Cruz
xhe89@ucsc.edu

Xin Eric Wang
UC Santa Cruz
xwang366@ucsc.edu

Abstract

Despite the success of Transformer models in vision and language tasks, they often learn knowledge from enormous data implicitly and cannot utilize structured input data directly. On the other hand, structured learning approaches such as graph neural networks (GNNs) that integrate prior information can barely compete with Transformer models. In this work, we aim to benefit from both worlds and propose a novel Multimodal Graph Transformer for question answering tasks that requires performing reasoning across multiple modalities. We introduce a graph-involved plug-and-play quasi-attention mechanism to incorporate multimodal graph information, acquired from text and visual data, to the vanilla self-attention as effective prior. In particular, we construct the text graph, dense region graph, and semantic graph to generate adjacency matrices, and then compose them with input vision and language features to perform downstream reasoning. Such a way of regularizing self-attention with graph information significantly improves the inferring ability and helps align features from different modalities. We validate the effectiveness of Multimodal Graph Transformer over its Transformer baselines on GQA, VQAv2, and MultiModalQA datasets.

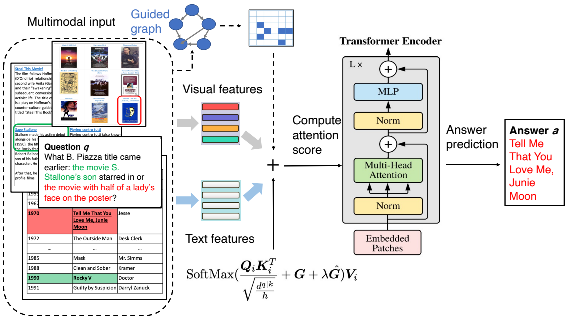

Multimodal input
Guided graph
Question *q*
What B. Piazza title came earlier: the movie S. Stallone's son starred in or the movie with half of a lady's face on the poster?
Visual features
Text features
Compute attention score
Transformer Encoder
L ×
MLP
Norm
Multi-Head Attention
Norm
Embedded Patches
Answer prediction
Answer *a*
Tell Me That You Love Me, Junie Moon
$\mathrm{SoftMax}(\frac{\boldsymbol{Q}_i\boldsymbol{K}_i^T}{\sqrt{\frac{d^{q|k}}{h}}}+\boldsymbol{G}+\lambda\hat{\boldsymbol{G}})\boldsymbol{V}_i$

Figure 1: Overview of Multimodal Graph Transformer. It takes visual features, text features, and their corresponding generated graphs as inputs. The generated graph is first converted to an adjacency matrix to induce the mask matrix $\boldsymbol{G}$. The modified quasi-attention score in the Transformer is computed to infer the answer. In the formular, $\boldsymbol{G}$ is the graph-induced matrix constructed by concatenating adjacency matrices both from the vision and the language end. $\hat{\boldsymbol{G}}$ is the trainable bias. The input features from different modalities are fused along with graph info to perform downstream reasoning.

1 Introduction

A myriad of complex real-world tasks require both prior knowledge and reasoning intelligence (Yi et al., 2018a; Ilievski and Feng, 2017). These days, vision-and-language reasoning tasks such as as vision question answering (VQA) (Antol et al., 2015) and multimodal question answering (MultiModalQA) (Talmor et al., 2021) post further needs for integrating structured info from different input modalities and thus perform reasoning. Towards this, two questions yield: What is the best way to integrate prior knowledge and reasoning components from multiple modalities in a single model? How would such an integration lead to accurate models, while being more computationally efficient and allowing for significantly more interpretability? Such questions are important to address when scaling reasoning systems to real-world use cases.

These years, there are a spectrum of methods in the literature exploring different ways of integrating structured prior information. Graph neural networks (GNNs) (Wu et al., 2020), have been widely used in representation learning on graphs. Some experts tried to investigate the embedding of the structured information by resorting to them. However, GNNs are inefficient (Wu et al., 2020) and they can barely compete with Transformer models. Besides, most GNNs are designed to learn node representations on fixed and homogeneous graphs. Thereby, it is suboptimal to operate GNNs on vision-and-language tasks such as visual question answering (VQA), where graphs encountered in these problems (e.g. scene graphs) can be more complex; Alternatively, knowledge graphs (KGs),

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 189–200
May 2-6, 2023 ©2023 Association for Computational Linguistics

such as Freebase (Bollacker et al., 2008), represent world-level factoid information of entities and their relations in a graph-based format, surfaced these years. They have been successfully used in vision and language applications including VQA (Marino et al., 2019). However, they have not been dedicated to be applied to our scenario, more concretely, we aim at filling the gap of capturing prior knowledge in Transformer models.

To mitigate deficiencies of the existing methods, this paper proposes a novel plug-and-play graph-involved Transformer-based method for multimodal question answering tasks. Our method is *Multimodal Graph Transformer* in the sense that it is built upon the well-established Transformer (Vaswani et al., 2017a) backbone, albeit with several key fundamental differences. First, we introduce a systematic scheme to convert text graphs, dense region graphs, and semantic graphs from vision and language tasks to adjacency matrices to use in our method. Second, instead of directly computing the attention score, we learn the newly proposed quasi-attention score with graph-induced adjacency matrices live at its heart, to signify the importance of learning relative importance as a highly effective inductive bias for computing the quasi-attention score. Third, different from previous Transformer methods, where self-attention are fully learned from data, we switch gears to introduce the graph-structured information in the self-attention computation to guide the training of Transformers as shown in Figure 1.

The main contributions are summarized below:

- We propose a novel Multimodal Graph Transformer learning framework that combines multimodal graph learning from unstructured data with Transformer models.
- We introduce a modular plug-and-play graph-involved quasi-attention mechanism with a trainable bias term to guide the information flow during training.
- The effectiveness of the proposed methods is empirically validated on GQA, VQA-v2, and MultiModalQA tasks.

2 Related Works

2.1 Multimodal question answering

Visual Question Answering (VQA)(Antol et al., 2015) has been a prominent topic in the field of multimodal question answering, garnering significant attention and advancing significantly since the introduction of the first large-scale VQA dataset byAntol et al. (2015). To answer VQA questions, models typically leverage variants of attention to obtain a representation of the image that is relevant to the question (Andreas et al., 2016; Yang et al., 2015; Xu and Saenko, 2016; Fukui et al., 2016; Lu et al., 2016). A plethora of works (Liang et al., 2021; Hudson and Manning, 2018; Yi et al., 2018b; Xiong et al., 2016; Kim et al., 2018; Teney et al., 2017a) have attempted to enhance the reasoning capability of VQA models, with Teney et al. (2017a) proposing to improve VQA using structured representations of the scene contents and questions. They developed a deep neural network that leverages the structure in these representations and builds graphs over scene objects and question words. The recent release of MultiModalQA (Talmor et al., 2021), a dataset that demands joint reasoning over texts, tables, and images, has received widespread attention. However, similar to VQA, existing MultiModalQA methods have not fully utilized structured information from the input concepts. To address this, we propose a combination of multimodal graph learning and Transformer models to improve question answering across inputs from multiple different modalities.

2.2 Attention mechanisms

The attention mechanism (Xu et al., 2015a,b; Devlin et al., 2018), has dramatically advanced the field of representation learning in machine learning. The attention mechanism is introduced in Vaswani et al. (2017b) and widely used in language tasks (i.e., abstract summarization (Xu et al., 2020)), machine translation (Bahdanau et al., 2014), reading comprehension (Dai et al., 2020), question answering (Min et al., 2019), etc. Zhang et al. (2020) proposes using syntax to guide the text modeling by incorporating explicit syntactic constraints into attention mechanisms. Meanwhile, it has seen increasing application in multimodal tasks (Li et al., 2020; Nam et al., 2017; Lu et al., 2016), where it is usually used for learning of interactions between multiple inputs. Following their success, multimodal Transformer models (Chen et al., 2019; Hu et al., 2020; Sun et al., 2019) have also shown impressive results on several vision-and-language tasks. Yun et al. (2019) proposes Graph Transformer Networks (GTNs) that can generate new graph structures and learn effective node representation on the new graphs in an end-to-end fashion. Different from these works, our work incorporates graph information from different modalities into the Transformer to improve the reasoning ability.

2.3 Exploiting graphs in multimodal reasoning

Considering that graph priors can transfer commonalities and mitigate the gap between visual and language domains, researchers explore how to use graphs (Teney et al., 2017b; Yu et al., 2020) properly in both tasks. In recent years, many classes of GNNs have been developed for both tasks which are divided into two approaches: spectral (Bruna et al., 2013) and non-spectral methods (Chen et al., 2018). Graphs can also be transferred into latent variables by GCN (Yang et al., 2019a; Yao et al., 2018), which can be directly utilized by models. However, the need for aligning graph priors from different modalities to do reasoning limits the use of graph priors. Our work addresses this problem via the graph-involved quasi-attention mechanism.

2.4 Pretraining

Pretrained models in computer vision (Simonyan and Zisserman, 2014; He et al., 2016) and NLP (Devlin et al., 2018; Yang et al., 2019b; Liu et al., 2019), have achieved state-of-the-art performances in many downstream tasks (Thongtan and Phienthrakul, 2019; White et al., 2017; Karpathy and Fei-Fei, 2015; Lee et al., 2018; Ren et al., 2015b). Other pretrained models such as VLBERT (Lu et al., 2019; Sun et al., 2019) and ViLT (Kim et al., 2021) also demonstrate their effectiveness on downstream vision-language tasks. Recent works on vision-language pretraining such as OSCAR (Li et al., 2020) perform cross-modal alignment in their visual-language pretraining models. Likewise, our proposed method includes cross-modality alignment, which is critical for reasoning. Our proposed modular plug-and-play graph-involved quasi-attention mechanism is also model-agnostic and can be also applied to other pretrained Transformer-based vision and language models.

3 Multimodal Graph Transformer

3.1 Background on Transformers

The Transformer layer (Vaswani et al., 2017b) consists of two modules: a multi-head attention and a feed-forward network (FFN). Specifically, each head is represented by four main matrices: the query matrix $\boldsymbol{W}_i^q \in \mathbb{R}^{d^m \times d^q/h}$, the key matrix $\boldsymbol{W}_i^k \in \mathbb{R}^{d^m \times \frac{d^k}{h}}$, the value matrix $\boldsymbol{W}_i^v \in \mathbb{R}^{d^m \times \frac{d^v}{h}}$, and the output matrix $\boldsymbol{W}_i^o \in \mathbb{R}^{\frac{d^v}{h} \times d^o}$, and takes the hidden states $\boldsymbol{H} \in \mathbb{R}^{l \times d^m}$ of the previous layer as input, where d denotes the dimension of the model, h represents the number of head, and i denotes the index of layer number. The output of attention is given by:

$$\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i = \boldsymbol{H}\boldsymbol{W}_i^q, \boldsymbol{H}\boldsymbol{W}_i^k, \boldsymbol{H}\boldsymbol{W}_i^v \tag{1}$$

$$\text{Attention}\left(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i\right) = \text{SoftMax}\left(\frac{\boldsymbol{Q}_i \boldsymbol{K}_i^T}{\sqrt{\frac{d^{q|k}}{h}}}\right)\boldsymbol{V}_i \tag{2}$$

$$\boldsymbol{H}_i = \text{Attention}\left(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i\right)\boldsymbol{W}_i^o \tag{3}$$

where $\boldsymbol{Q}_i \in \mathbb{R}^{l \times \frac{d^q}{h}}, \boldsymbol{K}_i \in \mathbb{R}^{l \times \frac{d^k}{h}}, \boldsymbol{V}_i \in \mathbb{R}^{l \times \frac{d^v}{h}}$ are obtained by the linear transformations of $\boldsymbol{W}_i^q, \boldsymbol{W}_i^k, \boldsymbol{W}_i^v$ respectively. $\text{Attention}(\cdot)$ is the scaled dot-product attention operation. Then output of each head is transformed to $\boldsymbol{H}_i \in \mathbb{R}^{l \times d^o}$ by $\boldsymbol{W}_i^o$.

3.2 Framework overview

The entire framework of the proposed Multimodal Graph Transformer method is depicted in Figure 2. Without loss of generality, we assume the end task is VQA in the following discussion while noting that our framework can be applied to other vision-language tasks, such as multimodal question answering.

Given the input images and questions, the framework first constructs three graphs, including the semantic graph, dense region graph, and text graph, which will be described in more detail in the following sections. The graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ represents the set of nodes in the graph and $\mathcal{E}$ represents the edges connecting them, is fed into Transformers to guide the training process.

3.3 Multimodal graph construction

We build three types of graphs and feed them into Transformers: *text graph*, *semantic graph*, and *dense region graph*. We now introduce them in detail.

Text graph The task of Visual Question Answering involves a combination of an image, a question, and its corresponding answer. To process the question, we extract the entities and create a text graph representation. We then build the graph $G = (\mathcal{V}, \mathcal{E})$

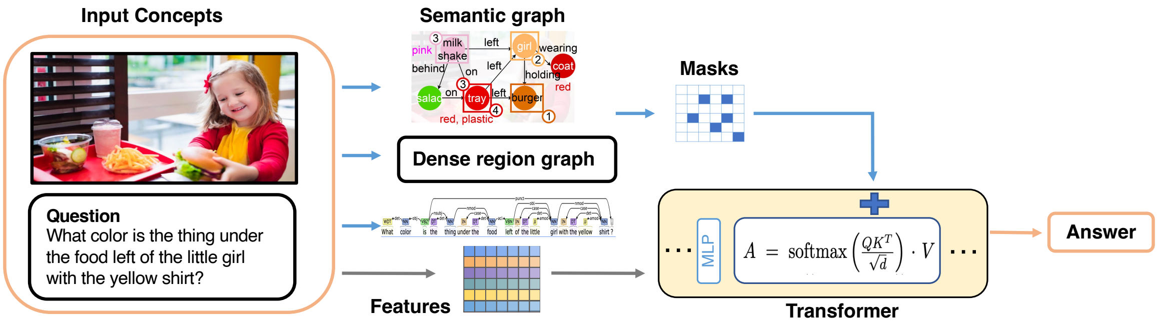


Figure 2: The figure illustrates the overall framework of our Multimodal Graph Transformer. The input from different modalities are processed and transformed into corresponding graphs, which are then converted into masks and combined with their features to be fed into Transformers for downstream reasoning. In detail, semantic graphs are created through scene graph generation methods, dense region graphs are extracted as densely connected graphs, and text graphs are generated through parsing.

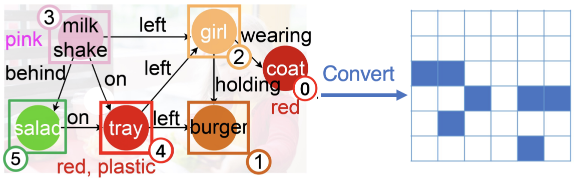


Figure 3: The naive demonstration of converting a semantic graph into an adjacency matrix. Cells in blue means ‘0’s for that element in the graph matrix, while white ones means ‘-inf’s. We employ the matrix as the mask when computing the quasi-attention.

as shown in the left of Figure 2. The set of nodes, $\mathcal{V}$, represents the entities and the set of edges, $\mathcal{E}$, represents the relationships between the pairs of entities. This results in:

- A set of N entities, each represented by a vector of token embeddings, that constitute the nodes of the graph.
- A set of pairwise relations between entities, forming the edges of the text graph. The relationship between entities i and j is represented by a vector e_{ij} which encodes the relative relationships.

Semantic graph In tasks such as multimodal question answering, there might be additional inputs in the form of tables or lengthy paragraph sentences. To handle these inputs, a linear representation of the table can be created and a semantic graph can be constructed using a similar approach. They are processed using the scene graph parser (Zhong et al., 2021), which transforms the text sentence into a graph of entities and relations, as depicted in Figure 3. The output of the scene graph parser includes:

- A set of N words that constitute the nodes of the semantic graph, where N is the number of parsed words in the texts.
- A set of possible pairwise relations between words, such as "left" and "on" as shown in Figure 3, which constitute the edges of our graph. An edge between words connecting j to i is represented by e_{ij}, namely, the connectivity is indicated as: $e_{ij} = \begin{cases} 0, & i,j \ \text{ not connected} \\ 1, & i,j \ \text{ connected} \end{cases}$.

Dense region graph The visual features are extracted by slicing the input images into patches and flattening them. A dense region graph $G = (\mathcal{V}, \mathcal{E})$ is then converted into masks, with $\mathcal{V}$ being the set of extracted visual features and $\mathcal{E}$ being the set of edges connecting each feature node, following the method described in (Kim et al., 2021). This results in a graph that is nearly fully connected.

The resulting three graphs are then transformed into adjacency matrices, where the elements are either -∞ or zero. The conversion process is depicted in Figure 3 using the semantic graph as an example. These adjacency matrices are used inside the scaled dot-product attention to control the flow of information, by masking out (setting to $-\infty$) the values.

3.4 Graph-involved quasi-attention

In order to effectively utilize structured graph knowledge in our self-attention computation, we incorporate the graph as an extra constraint in each

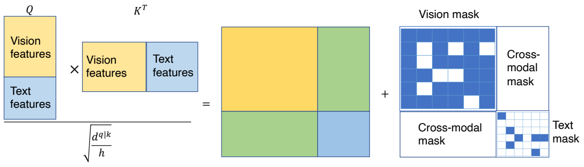


Figure 4: A naive demonstration of adding the graph-induced mask while computing the quasi-attention when the inputs are from two modalities. The visual mask is the mask converted from the dense region graph and the text mask is converted from the text graph. The cross-modal mask, which is always set as an all-zero matrix, is imposed to encourage the model to learn the cross-attention between the image features and text features, thus facilitating the alignment across them.

attention head by converting it into an adjacency matrix. The graph matrix, denoted as $\boldsymbol{G}$, is constructed by combining various masks. An illustration of this process can be seen in Figure 4. The visual mask is generated from the dense region graph, while the text mask is derived from the text graph. Additionally, the cross-modal mask is set to an all-zero matrix to encourage the model to learn the cross-attention between visual and text features, thereby promoting alignment across the different modalities.

Within the context of adding graph information, when vision graph mask and text graph mask are concatenated and aligned with image and text features, we believe that a more flexible masking-out mechanism is beneficial, rather than keeping a single constant mask matrix inside the Softmax operation. Drawing insights from Liu et al. (2021), where they include a relative position bias to each head in computing similarity, we also intuitively parameterize a trainable bias $\hat{\boldsymbol{G}}$ and involve it in the training process. Finally, we compute the quasi-attention as follows:

$$\text{Attention} = \text{SoftMax}(\frac{\boldsymbol{Q}_i\boldsymbol{K}_i^T}{\sqrt{\frac{d^{q|k}}{h}}} + \boldsymbol{G} + \lambda\hat{\boldsymbol{G}})\boldsymbol{V}_i, \tag{4}$$

where λ is the tradeoff hyper-parameter that controls the contribution of $\hat{\boldsymbol{G}}$, and $\boldsymbol{G}$ is our graph-induced matrix constructed by concatenating a graph matrix both from the vision and the language end. Here for clear clarification, we use $\boldsymbol{G}$ and $\hat{\boldsymbol{G}}$ to distinguish the graph matrices fixed and trainable, respectively. During training, $\boldsymbol{G}$ is frozen as before and does not receive gradient updates, while $\hat{\boldsymbol{G}}$ contains trainable parameters.

We now introduce the motivation behind adding two types of graph matrices. We perform the masking process by adding $\boldsymbol{G}$ when computing the quasi-attention because it can be interpreted as a form of attentional pooling (learning to align), in which each element of $\boldsymbol{G}$ pools all relevant information across all elements of the relative importance matrix computed by $\left(\frac{\boldsymbol{Q}_i\boldsymbol{K}_i^T}{\sqrt{\frac{d^{q|k}}{h}}}\right)$. Hence during fine-tuning, the model ignores redundant features and only focuses on useful information. The mask can also force the model to learn the cross attention between features from the images and questions and perform aligning across them. And the trainable bias $\hat{\boldsymbol{G}}$ captures information gained during the training process. Such information is valuable for fine-tuning, making the Transformer more robust and helping it gain numerical stability.

3.5 Training

The interdependence of output features from various modalities calls for a unified optimization approach for the Transformers in both the visual question answering and multimodal question answering tasks. To accomplish this, we implement a kind of end-to-end training, which ensures the optimality of the models. The final outcome of our models is a classification logit, which is generated by the VQA models that select the best answer from the available candidate answers. To evaluate the accuracy of the models, we compute the cross-entropy loss (Zhang and Sabuncu, 2018) using the output logits produced by the Transformer. This measure helps us determine the difference between the predicted class probabilities and the actual class labels.

4 Experiments

4.1 Datasets

VQA v2 The VQA v2 dataset (Goyal et al., 2017) extends the VQA (Antol et al., 2015) dataset to better balance visual and textual information through the collection of complementary images. Each question in VQA v2 is associated with a pair of similar images with different answers, resulting in a total of 1.1 million QA pairs and 204,000 images. The data split for VQA v2 includes a training set with 83,000 images and 444,000 questions, a validation set with 41,000 images and 214,000 questions, and a test set with 81,000 images and 448,000 questions. The annotated answers are in natural language, but they are commonly converted to a classification task with 3,129 answer classes.

Table 1: Accuracy (%) comparison of different methods on the GQA and VQA v2 test-dev. Ours has the second best performance and is comparable to state-of-the-art methods. After applying our proposed quasi-attention mechanism and exploiting the use of graphs, there is also a 2% improvement of overall accuracy on the LXMERT baseline, suggesting the generalization ability of our method.

Dataset	Method	Open questions	Binary questions	Overall accuracy
GQA	LXMERT (Tan and Bansal, 2019)	-	-	60.0
	LXMERT w/ Graph (Tan and Bansal, 2019)	-	-	61.4
	HANs (Kim et al., 2020)	-	-	69.4
	NSM (Hudson and Manning, 2019b)	49.3	78.9	63.2
	OSCAR (Li et al., 2020)	-	-	61.6
	VinVL (Zhang et al., 2021)	-	-	65.1
	Multimodal Graph Transformer (Ours)	59.4	80.5	68.7
VQA v2	LXMERT (Tan and Bansal, 2019)	-	-	72.4
	HANs (Kim et al., 2020)	-	-	65.1
	NSM (Hudson and Manning, 2019b)	-	-	63.0
	OSCAR (Li et al., 2020)	-	-	73.8
	VinVL (Zhang et al., 2021)	-	-	76.6
	Multimodal Graph Transformer (Ours)	66.5	87.0	74.5

As described by Anderson et al. (2018), the model selects the answer to each question from a set of 3,129 most frequent answers. Following this convention, we fine-tune the multimodal graph transformer model on the VQAv2 training and validation sets, while reserving 1,000 validation images and related questions for internal validation.

GQA The GQA dataset contains 22M questions over 113K images. The questions in GQA are designed to require multi-hop reasoning to test the reasoning skills of VQA models. GQA greatly increases the complexity of the semantic structure of questions, leading to a more diverse function set. The real-world images in GQA also bring in a bigger challenge in visual understanding. We conduct experiments on the public splits (Hudson and Manning, 2019a) of the GQA dataset and also treat the task as the classification task reffering to the VQA v2 setting.

MultiModalQA MultiModalQA (MMQA) contains 29, 918 questions. We split the dataset into 23,817 training, 2,441 development (dev.), and 3,660 test set examples referring to the official split. Around 60% of the questions in MMQA are compositional. The answer for each question can be a single answer or a list of answers.

4.2 Baselines

We compare with four state-of-the-art VQA models: LXMERT (Tan and Bansal, 2019), NSM (Hudson and Manning, 2019b), OSCAR (Li et al., 2020), and VinVL (Zhang et al., 2021).

- LXMERT (Tan and Bansal, 2019) designs five pretraining tasks: masked language modeling, feature regression, label classification, cross-modal matching, and image question answering to pretrain a large Transformer model. Towards this, a large-scale Transformer (Vaswani et al., 2017b) model is built that consists of three encoders: an object relationship encoder, a language encoder, and a cross-modal encoder.

- NSM (Hudson and Manning, 2019b) predicts a probabilistic graph that represents its underlying semantics and performs sequential reasoning over the graph to traversing its nodes to make the inference.

- OSCAR (Li et al., 2020) uses object tags detected in images as anchor points to significantly ease the learning of alignments, improving previous methods and using self-attention to learn image-text semantic alignments.

- VinVL (Zhang et al., 2021) developed a new object detection model to create better visual features of images than previous classical object detection models.

We compare with four baselines introduced in the MultiModalQA paper (Talmor et al., 2021): Question-only (Kaushik and Lipton, 2018), Context-only (Kaushik and Lipton, 2018), AutoRouting, ImplicitDecomp.

- Question-only is a sequence-to-sequence model that directly generates the answer given the question.

- Context-only first predicts the question type using the classifier and then feed in the relevant context to predict the answer.

- AutoRouting first determines the modality where the answer is expected to occur, and then runs the corresponding single-modality module.

- ImplicitDecomp is a 2-hop implicit decomposition baseline and so far the state-of-the-art method on the MultiModalQA dataset.

4.3 Implementation details

The input texts undergo preprocessing using a scene graph parser which extracts entities and their relationships. The text features are obtained through a pre-trained BERT tokenizer, allowing us to extract text spans of individual entities and text spans containing two related entities. As for images, we employ the methods described in Dosovitskiy et al. (2020); Kim et al. (2021) to extract visual features and create graph masks. This involves resizing the shorter edge of the input images while preserving the aspect ratio and limiting the longer edge, followed by patch projection and padding for batch training. The resulting patch embeddings are used as inputs along with constructed dense region graph that is densely connected. The Transformer backbone used in this setting is the pretrained VIT-B-32 (Dosovitskiy et al., 2020) version, consisting of 12 layers with a hidden size of $H = 768$, layer depth of $D = 12$, patch size of $P = 32$, a multi-layer perceptron size of 3072, and 12 attention heads. To test this setting, all inputs and graphs are merged and processed by the Transformer backbone, which learns from features from different modalities.

4.3.1 MultiModalQA

We further investigate the effectiveness of our proposed method on MultiModalQA (Talmor et al., 2021), a recently introduced and demanding task that requires joint reasoning across various modalities such as texts, images, tables, etc. We employ a Multimodal Graph Transformer to tackle the task, using the same approach for extracting vision and text features as in VQA. Additional modalities, such as tables, are encoded by linearizing them and utilizing pre-trained models like RoBERTa-large (Liu et al., 2019). After generating text graphs, semantic graphs, and dense region graphs from input questions, text, tables, and images, we feed them along with the extracted features into the Transformer. Unlike the Transformer used in VQA, which takes inputs from two modalities, the MultiModalQA Transformer accepts input from three modalities and performs the final reasoning.

Table 2: EM (%) and F1 (%) of Multimodal Graph Transformer and its Transformer baseline on questions in MultiModalQA that require reasoning over multiple modalities. Incorporating graph information into the Multimodal Graph Transformer can boost about 2% F1 and 4% EM performance.

Method	EM	F1
Question-only	16.9	19.5
Context-only	6.6	8.5
AutoRouting	32.0	38.2
ImplicitDecomp	46.5	51.7
Human	84.8	90.1
Multimodal Transformer w/o Graph	50.1	56.4
Multimodal Graph Transformer (Ours)	52.1	57.7

4.4 Results and analysis

Table 1 presents a comparison of the accuracy of our proposed method on the GQA dataset with previous state-of-the-art methods. Our proposed method ranks second in terms of accuracy and outperforms the third best method by a substantial margin, with an absolute improvement of over 3% in overall accuracy. The performance of our method is comparable to the state-of-the-art method.

We also conducted experiments on the VQA v2 dataset, and the results are summarized in Table 1 and Table 3. As shown, there are significant improvements over methods without graphs, suggesting that incorporating graph information into the Transformer is effective.

Additionally, after incorporating our proposed graph method into LXMERT, we can observe a boost in overall accuracy on the GQA dataset, demonstrating the generalization ability of the proposed method in incorporating graph information into quasi-attention computation.

Table 2 compares the Exact Match (EM) and average F1 score of our proposed method on the MultiModalQA dataset with the baseline. The results show that our proposed method outperforms the baseline without the aid of graph information, demonstrating the generalization of our method to more complicated vision-and-language reasoning

Table 3: Ablation Studies on the GQA and VQA v2 Validation Sets. The figure demonstrates the effectiveness of incorporating graph information into the Transformer architecture through ablation studies performed on the GQA and VQA v2 validation sets. The results of these studies clearly indicate that including graph information can lead to an improvement in performance.

Dataset	Method	Open questions	Binary questions	Overall accuracy
GQA	One-modality Transformer	47.7	78.1	62.7
	Multimodal Transformer w/o Graph	49.9	81.0	65.4
	Ours	**60.1**	**90.2**	**72.4**
VQA v2	One-modality Transformer w/ one Transformer	60.5	85.4	70.1
	Multimodal Transformer w/o Graph	64.8	86.3	72.1
	Ours	**66.7**	**87.2**	**74.6**

tasks.

4.5 Ablation studies

We perform ablation studies to verify the necessity of using two-stream inputs with the help of graphs to deal with input from different modalities, with GQA dataset as our testing bed. For all experiments, we use the overall accuracy as the evaluation metric.

The results presented in Table 3 show the superiority of our proposed Multimodal Graph Transformer over the method where a single modality input is fed into a Transformer. Our method, which involves dividing the input streams into two separate parts and processing each part through a Transformer, outperforms the Multimodal Transformer without Graph. This demonstrates the beneficial effect of incorporating graph information into the processing of the input data and performing training. The use of different input features with the help of graphs allows for a better alignment of the information from different modalities, which is reflected in the improved performance of our proposed method.

4.6 Qualitative results

One qualitative example is shown in Figure 5. As can be seen, predictions from Multimodal Graph Transformer are more relevant to contents of the input image as the graph information improves the inferring ability of the Transformer, which further indicates the effectiveness of Multimodal Graph Transformer.

5 Conclusions

In this paper, we have presented a novel method to integrate structured graph information to guide the Transformers training. Our method can model interactions between different modalities and achieves competitive performance on multimodal reasoning tasks such as VQA and MultiModalQA. Experimental results show that our method outperforms many other methods on the GQA dataset. More importantly, the proposed quasi-attention mechanism is model-agnostic and it is possible to apply it to other Transformer-based methods. We will test our methods on other vision-and-language reasoning tasks and include the comparison with existing graph representation learning methods in our future work.

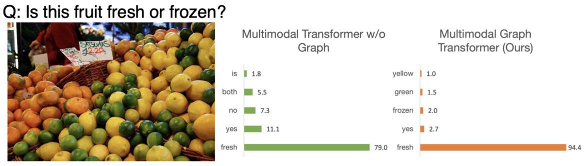


Figure 5: A qualitative comparison from VQA v2. *fresh* is the ground truth. Predictions from the Multimodal Graph Transformer (ours) are more relevant to the contents of the input image and achieve a higher confidence score over the ground truth.

6 Limitations and Potential Risks

The Limitations of the proposed Multimodal Graph Transformer include the potential preservation of fairness and bias issues inherent in the pretrained Transformer models, despite the involvement of graph information. Additionally, the integration of graphs may introduce new biases that can further exacerbate the problem. One potential source of bias is the vision-and-language dataset itself, which may favor majority cases and overlook minority cases. Unfortunately, the proposed method is not equipped to address these biases and issues, making further research and consideration crucial when building upon or directly using this method for vision and language tasks.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48.

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

Stanislaw Antol, C Lawrence Zitnick, and Devi Parikh. 2014. Zero-shot learning via visual abstraction. In *ECCV*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, page 1247–1250, New York, NY, USA. Association for Computing Machinery.

Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. 2013. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*.

Jie Chen, Tengfei Ma, and Cao Xiao. 2018. Fastgcn: fast learning with graph convolutional networks via importance sampling. *arXiv preprint arXiv:1801.10247*.

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

Zihang Dai, Guokun Lai, Yiming Yang, and Quoc V Le. 2020. Funnel-transformer: Filtering out sequential redundancy for efficient language processing. *arXiv preprint arXiv:2006.03236*.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*.

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

Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. 2020. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9992–10002.

Drew A Hudson and Christopher D Manning. 2018. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067*.

Drew A Hudson and Christopher D Manning. 2019a. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709.

Drew A Hudson and Christopher D Manning. 2019b. Learning by abstraction: The neural state machine. *arXiv preprint arXiv:1907.03950*.

Ilija Ilievski and Jiashi Feng. 2017. Multimodal learning and reasoning for visual question answering. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 551–562.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Divyansh Kaushik and Zachary C Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. *arXiv preprint arXiv:1808.04926*.

Eun-Sol Kim, Woo Young Kang, Kyoung-Woon On, Yu-Jung Heo, and Byoung-Tak Zhang. 2020. Hypergraph attention networks for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14581–14590.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *NIPS*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216.

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

Weixin Liang, Yanhao Jiang, and Zixuan Liu. 2021. Graghvqa: Language-guided graph neural networks for graph-based visual question answering. *arXiv preprint arXiv:2104.10283*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*.

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

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29:289–297.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3195–3204.

Sewon Min, Danqi Chen, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019. Knowledge guided text retrieval and reading for open domain question answering. *arXiv preprint arXiv:1911.03868*.

Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 299–307.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015a. Image question answering: A visual semantic embedding model and a new dataset. *NIPS*.

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

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor segmentation and support inference from rgbd images. In *ECCV*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7464–7473.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodalqa: Complex question answering over text, tables and images. *arXiv preprint arXiv:2104.06039*.

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

Damien Teney, Lingqiao Liu, and Anton van Den Hengel. 2017a. Graph-structured representations for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Damien Teney, Lingqiao Liu, and Anton van Den Hengel. 2017b. Graph-structured representations for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Tan Thongtan and Tanasanee Phienthrakul. 2019. Sentiment classification using document embeddings trained with cosine similarity. In *ACL SRW*.

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

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

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24.

Caiming Xiong, Stephen Merity, and Richard Socher. 2016. Dynamic memory networks for visual and textual question answering. In *International conference on machine learning*, pages 2397–2406. PMLR.

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

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015a. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015b. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*.

Song Xu, Haoran Li, Peng Yuan, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. Self-attention guided copy mechanism for abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1355–1362.

Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019a. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10685–10694.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019b. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. 2015. Stacked attention networks for image question answering. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21–29.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 684–699.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B Tenenbaum. 2018a. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *arXiv preprint arXiv:1810.02338*.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B Tenenbaum. 2018b. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *arXiv preprint arXiv:1810.02338*.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*, 1:12.

Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. 2019. Graph transformer networks. *Advances in Neural Information Processing Systems*, 32:11983–11993.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588.

Zhilu Zhang and Mert R Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. In *32nd Conference on Neural Information Processing Systems (NeurIPS)*.

Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, Hai Zhao, and Rui Wang. 2020. Sg-net: Syntax-guided machine reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9636–9643.

Yiwu Zhong, Jing Shi, Jianwei Yang, Chenliang Xu, and Yin Li. 2021. Learning to generate scene graph from natural language supervision. *arXiv preprint arXiv:2109.02227*.

C Lawrence Zitnick and Devi Parikh. 2013. Bringing semantics into focus using visual abstraction. In *CVPR*.

A Appendix

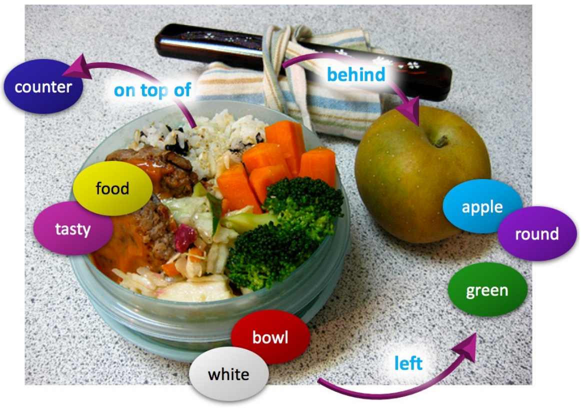
A.1 Visual Question Answering dataset

To address the problem of visual question answering, a number of visual question answering datasets have been developed. The comparison of them is shown in Table 4. The VQA dataset (Antol et al., 2015) is developed on real images in MS COCO (Lin et al., 2014) and abstract scene images

Table 4: Comparison of VQA datasets

	Source of images	# images	# QA pairs	Answer type	Evaluation metrics
DAQUAR	NYU-Depth V2	1,449	12,468	Open	Accuracy&WUPS
VQA	COCO	204K	614K	Open/MC	Accuracy
VQA v2	COCO	204K	1.1M	Open/MC	Accuracy
COCO-QA	COCO	123K	118K	Open/MC	Accuracy
CLEVR	Generated	100K	999K	Open	Accuracy
GQA	Visual Genome	113K	22M	Open	Accuracy

in Antol et al. (2014); Zitnick and Parikh (2013). The question-answer pairs are created by human annotators who are encouraged to ask “interesting" and “diverse" questions. VQA v2 (Goyal et al., 2017) is extended from the VQA (Antol et al., 2015) dataset to achieve more balance between visual and textual information by collecting complementary images in a way that each question is associated with a pair of similar images with different answers; In the COCO-QA (Ren et al., 2015a) dataset, the question-answer pairs are automatically generated from image captions based on syntactic parsing and linguistic rules; DAQUAR (Malinowski and Fritz, 2014) is built on top of the NYU-Depth V2 dataset (Silberman et al., 2012) which contains RGBD images of indoor scenes. DAQUAR consists of (1) synthetic question-answer pairs that are automatically generated based on textual templates and (2) human-created question-answer pairs produced by five annotators; CLEVR (Johnson et al., 2017) is a dataset developed on rendered images of spatially related objects (including cube, sphere, and cylinder) with different sizes, materials, and colors. The locations and attributes of objects are annotated for each image. The questions are automatically generated from the annotations; GQA is a new dataset for real-world visual reasoning and compositional question answering, seeking to address key shortcomings of previous VQA datasets. Considering questions in GQA are most objective, unambiguous, compositional, and can be answered by reasoning only on the visual content. We mainly use the GQA dataset in this work as it best fits our goal of reasoning. We also evaluate our methods on the VQA v2 dataset as it is the most common and general VQA dataset so far.



*Is the **bowl** to the right of the **green apple**?*
*What type of **fruit** in the **image** is **round**?*
*What color is the **fruit** on the right side, red or **green**?*
*Is there any **milk** in the **bowl** to the left of the **apple**?*

Figure 6: Examples from the GQA dataset for visual reasoning and compositional question answering.

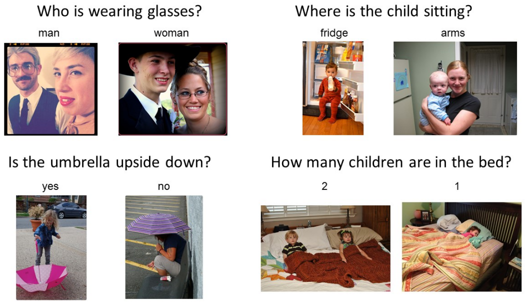

Who is wearing glasses?
man
woman
Where is the child sitting?
fridge
arms
Is the umbrella upside down?
yes
no
How many children are in the bed?
2
1

Figure 7: Examples from the VQA v2 dataset for Visual Question Answering.