Object Attribute Matters in Visual Question Answering

Peize Li[1*], **Qingyi Si**[2, 3*], **Peng Fu**[2, 3†], **Zheng Lin**[2, 3], **Yan Wang**[1, 4†]
[1]School of Artificial Intelligence, Jilin University, Changchun, China
[2]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[3]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[4]Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun, China
lipz21@mails.jlu.edu.cn, {siqingyi, fupeng, linzheng}@iie.ac.cn, wy6868@jlu.edu.cn

Abstract

Visual question answering is a multimodal task that requires the joint comprehension of visual and textual information. However, integrating visual and textual semantics solely through attention layers is insufficient to comprehensively understand and align information from both modalities. Intuitively, object attributes can naturally serve as a bridge to unify them, which has been overlooked in previous research. In this paper, we propose a novel VQA approach from the perspective of utilizing object attribute, aiming to achieve better object-level visual-language alignment and multimodal scene understanding. Specifically, we design an attribute fusion module and a contrastive knowledge distillation module. The attribute fusion module constructs a multimodal graph neural network to fuse attributes and visual features through message passing. The enhanced object-level visual features contribute to solving fine-grained problem like counting-question. The better object-level visual-language alignment aids in understanding multimodal scenes, thereby improving the model's robustness. Furthermore, to augment scene understanding and the out-of-distribution performance, the contrastive knowledge distillation module introduces a series of implicit knowledge. We distill knowledge into attributes through contrastive loss, which further strengthens the representation learning of attribute features and facilitates visual-language alignment. Intensive experiments on six datasets, COCO-QA, VQAv2, VQA-CPv2, VQA-CPv1, VQAvs and TDIUC, show the superiority of the proposed method.

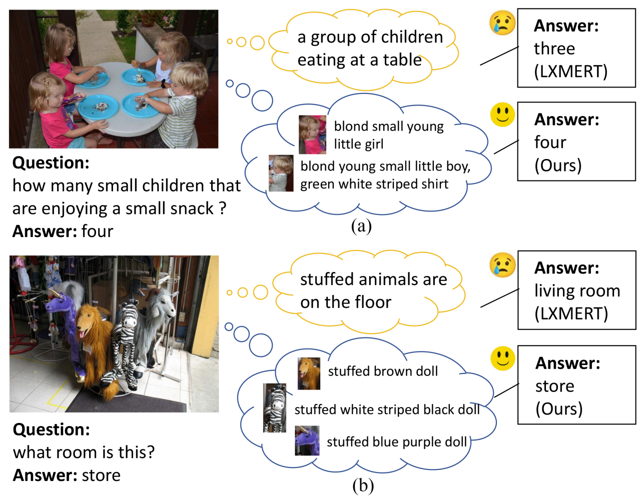


Figure 1: An illustration of our motivation. Compared with previous multimodal content, object-level attributes are indispensable in both object counting (a) and scene understanding (b).

Introduction

Visual Question Answering is a multimodal task involving the interaction of vision and language, which aims to answer the question based on visual image content. Most of the existing solutions (Kim, Jun, and Zhang 2018; Anderson et al. 2018; Li et al. 2019; Yu et al. 2019; Peng et al. 2022a; Si et al. 2023b) depend on visual relations, attention mechanisms and external knowledge to connect question information and associated visual clues. Visual relations (Li et al. 2019; Peng et al. 2022a) provide semantic connections and relative positions between the objects, aiding in enhancing the spatial understanding of image content. Attention mechanisms (Kim, Jun, and Zhang 2018; Anderson et al. 2018; Yu et al. 2019) give co-occurrence information in multimodal scenes, which enables the model to concentrate on the important words and visual elements. External knowledge (Gao et al. 2022; Gui et al. 2022) offers relevant background and topological relationships among the entities, which contribute to understanding the contextual information of multimodal scenes. However, both of these lack attributes of visual objects, which can directly offer fine-grained semantic information about visual objects. The object attributes cover a wide range of advanced concepts, including objects, scenes, actions and modifiers, which are indispensable for enhancing the understanding of object-level visual content and achieving object-level visual-language alignment. Next, we can better explain this idea through the following two examples.

An example from COCO-QA (Ren, Kiros, and Zemel 2015) dataset is shown in Figure 1(a). Answering this question first requires understanding the various types of children

[*]These authors contributed equally.
[†]Corresponding Authors.

in the image and then calculating the number of children. Object attributes provide different descriptive information for each child, which improves the model's ability to solve object-level fine-grained problem, such as enhanced counting ability. Another example from VQA-CPv2 (Agrawal et al. 2018) dataset is shown in Figure 1(b). Answering this question requires combining the semantic information of multiple objects in the scene, and then the model makes a comprehensive judgment. Object attributes improve the model's ability to solve complicated scene-understanding problem, which boosts the out-of-domain (OOD) performance. In summary, visual object attributes achieve object-level visual-language alignment, especially beneficial for the above two problems. Therefore, **O**bject **A**ttribute **M**atters in **V**isual **Q**uestion **A**nswering (OAM-VQA).

Recently, several methods have been well-developed to enhance VQA models using object attributes. Some prompt-based learning methods (Gui et al. 2022; Gao et al. 2022; Si et al. 2023b) utilize attributes to design prompts, other methods (Agrawal et al. 2018; Anderson et al. 2018; Nguyen et al. 2022) fuse attributes based on attention mechanisms. However, none of them achieve strong object-level visual-language alignment. As a result, they perform poorly on object-level fine-grained problem as well as complicated scene-understanding problem.

To address the aforementioned problem, we utilize object attributes to explicitly align visual and linguistic semantics. Specifically, our approach primarily consists of the Attribute Fusion Module (AFM) and the Contrastive Knowledge Distillation Module (CKDM). Attribute Fusion Module establishs a novel multimodal graph neural network to fuse the visual features and object attributes. Through updating nodes, the multimodal graph neural network iteratively aggregates information from neighboring nodes to capture detailed global information encompassing all objects. This allows the Attribute Fusion Module to learn both the shared characteristics among all objects and their individual attributes. In this way, the advanced object-level visual features contribute to addressing object-level fine-grained problem.

Contrastive Knowledge Distillation Module further enriches the representation of attribute features. Following TwO (Si et al. 2023b), this module firstly uses prompt to introduce a series of implicit knowledge stored in the visual-language pre-trained (VLP) models OFA (Wang et al. 2022), BLIP (Li et al. 2022) and BLIP2 (Li et al. 2023a). Then, it employs an enhanced transformer to encode knowledge. Through contrastive loss, we distill knowledge into attributes, which enhances the understanding of scenes and the model's robustness. Therefore, this module contributes to addressing complicated scene-understanding problem.

In summary, this paper explores the role of object attributes in visual question answering, and finds that object attributes are beneficial for enhancing the understanding of object-level visual content and facilitating the alignment between object-level visual and linguistic elements. The main contributions of this work contain:

- We propose a novel and effective method OAM-VQA that leverages object attributes to explicitly unify the visual and linguistic semantics.
- We design an Attribute Fusion Module and a Contrastive Knowledge Distillation Module, which respectively contribute to addressing object-level fine-grained problem and complicated scene-understanding problem.
- Extensive experimental results on six datasets, including COCO-QA, VQAv2, VQA-CPv2, VQA-CPv1, VQAvs and TDIUC, validate the effectiveness and generality of our approach.

Related Work

Incorporating Object Attribute in VQA. Recently, some inspiring works (Si et al. 2023b; Gui et al. 2022; Gao et al. 2022; Anderson et al. 2018; Nguyen et al. 2022) attempt to incorporate object attributes to address the VQA task and achieve remarkable progress. UpDn (Anderson et al. 2018) and VinVL (Zhang et al. 2021) directly leverage object attributes as input to learn effective visual representations. Different from focusing on enhancing the object detector, CFR-VQA (Nguyen et al. 2022) designs an elaborate BAN (Kim, Jun, and Zhang 2018) to fuse attribute features. However, it also unavoidably introduces some noise or ambiguous attribute information. The prompt-based learning methods (Si et al. 2023b; Gui et al. 2022; Gao et al. 2022) utilize attributes to obtain external knowledge from VLP models. These methods excel in leveraging broader cross-domain knowledge to solve the VQA task. However, they fail to achieve object-level visual-language alignment, which could lack the capability to address object-level fine-grained problem and scene-level understanding problem. Our method goes further in both directions: On the one hand, we establish the multimodal graph neural network to fuse object attributes. On the other hand, we do not merely introduce a series of knowledge. Furthermore, we effectively utilize knowledge to enrich attribute feature representation and promote object-level visual-linguistic alignment.

In recent years, numerous studies (Si et al. 2022b; Goyal et al. 2017; Ren, Kiros, and Zemel 2015; Si et al. 2023a) propose diverse VQA tasks to evaluate different types of core skills for addressing the visual question answering. One type of dataset focuses on image content understanding, such as COCO-QA (Ren, Kiros, and Zemel 2015) and TDIUC (Kafle and Kanan 2017). COCO-QA (Ren, Kiros, and Zemel 2015) is automatically generated based on image captions and can be classified into four main types: object, color, number and location. TDIUC (Kafle and Kanan 2017) is a task-driven image understanding dataset, where the questions can be categorized into 12 classes, such as counting and sentiment understanding. The OOD datasets have a notable difference in answer distribution between the training and testing sets, and the models that only learn biases from the training set struggle to perform well on OOD datasets. Common OOD datasets consist of VQA-CPv1/2 (Agrawal et al. 2018), VQAv2 (Goyal et al. 2017) and VQAvs (Si et al. 2022b). They are proposed for studying language bias problem. Furthermore, VQAvs (Si et al. 2022b) is a comprehensive dataset containing visual bias, language bias and multi-modal bias. We validate our approach on multiple datasets, which cover two important settings: image content understanding and out-of-distribution robustness. In this way, the visual question answering ability of the model is comprehensively assessed.

Graph Neural Network. Graph neural network (GNN) (Li and Moens 2022; Scarselli et al. 2008; Li et al. 2019; Gao et al. 2020; Zhu et al. 2020) is a highly effective framework for representing graph-structured data. GNNs follow the message passing scheme that updates each node's feature using its neighborhoods of nodes to capture specific patterns of a graph. Some encouraging works (Li and Moens 2022; Li et al. 2019; Gao et al. 2020; Zhu et al. 2020) study graph neural networks to solve the VQA task. For example, ReGAT (Li et al. 2019) represents the image as a graph and captures interactions between objects through the graph attention mechanism. Moreover, Mucko (Zhu et al. 2020) constructs a multimodal heterogeneous graph consisting of visual features, image captions and factual knowledge. It utilizes graph convolutional networks to capture multi-layer graph representations to predict the answers. Unlike the aforementioned approaches that update nodes based on modality-specific information, we establish a multimodal graph consisting of a visual sub-graph and an attribute sub-graph. Our approach updates node representation from interactions across different modalities to learn comprehensive attribute feature representations and better achieve object-level visual-linguistic alignment.

Methodology

In this section, we elaborate on the proposed OAM-VQA approach for visual question answering. Figure 2 shows OAM-VQA's overview, which contains: multimodal encoding, visual description module, attribute fusion module, contrastive knowledge distillation module and answer prediction module.

Multimodal Encoding

A multimodal encoder is used to encode question and image features. Most existing VQA models consider VQA as a multi-class classification task. Among them, LXMERT (Tan and Bansal 2019) is a transformer-based approach like BERT (Devlin et al. 2019), and can encode visual objects and question text into visual features and textual features. Besides, LXMERT is the most popular pre-trained visual-language model, thus we adopt it as the multimodal encoder in our proposed approach. Concretely, given a VQA dataset $\mathcal{D} = \{(\boldsymbol{I}_i, \boldsymbol{Q}_i, \boldsymbol{A}_i)\}_{i=1}^{N}$ with N samples, where $\boldsymbol{I}_i$, $\boldsymbol{Q}_i$ and $\boldsymbol{A}_i$ are the image, question and ground-truth answer of i-th sample respectively. LXMERT encodes image $\boldsymbol{I}_i$ and question $\boldsymbol{Q}_i$ separately in two streams, and extracts visual features $\boldsymbol{V}_i = \{\boldsymbol{o}_{i1}, \boldsymbol{o}_{i2}, ..., \boldsymbol{o}_{ij}\}_{j=1}^{M}$ and question features $\boldsymbol{T}_i$. M is the number of visual objects detected by Faster RCNN (Ren et al. 2015). The visual features $\boldsymbol{V}_i$ will serve as the initial representation in the attribute fusion module, and question features $\boldsymbol{T}_i$ will be further utilized in the contrastive knowledge distillation module.

Visual Description Module

Visual descriptions provide more descriptive semantic information about visual images, which effectively reduces the semantic gap between the two modalities. Given an image $\boldsymbol{I}_i$, following TwO (Si et al. 2023b), visual description module generates descriptive text at different levels, consisting of object-level attributes, image-level global captions and image-level detailed descriptions. First, we use the VinVL detector (Zhang et al. 2021) to obtain object-level attributes and utilize 300-dimensional Glove (Pennington, Socher, and Manning 2014) to acquire their word embeddings as the initial attribute features $\boldsymbol{E}_i = \{\boldsymbol{e}_{i1}, \boldsymbol{e}_{i2}, ..., \boldsymbol{e}_{ik}\}_{k=1}^{L}$. L is the number of attribute words. Then, we adopt the SOTA visual-language pretrained model BLIP2 (Li et al. 2023a) to generate image-level global captions and obtain the corresponding encoded features $\boldsymbol{C}_i$ in the same way. Besides, we apply a multimodal large language model mPLUG-Owl (Ye et al. 2023) to generate image-level detailed descriptions. mPLUG-Owl (Ye et al. 2023) is a multimodal model based on large language model. It has stronger language generation capabilities and is capable of generating descriptions more detailed than traditional image captions. And in the ablation experiments, we compare it with object-level attributes and image-level captions to explore their effects in VQA. The aforementioned descriptions of visual content will serve as the initial features for the subsequent attribute fusion module.

Attribute Fusion Module

Attribute fusion module guides information passing between the visual graph and the semantic graph. The goal of this module is to fuse object-level attributes and visual features to achieve better object-level visual-linguistic alignment.

Multimodal graph construction. Given an image $\boldsymbol{I}_i$, we first construct a multimodal graph composed of two fully connected sub-graphs, i.e., visual graph $\boldsymbol{G}_{iv}$ and semantic graph $\boldsymbol{G}_{it}$ for representing two modalities of information. In the visual graph $\boldsymbol{G}_{iv}$, each node $\boldsymbol{v}_{ij}$ represents each visual object. The initial representation $\boldsymbol{v}_{ij}^{(0)}$ is obtained through multimodal encoding. We set $\boldsymbol{v}_{ij}^{(0)} = \boldsymbol{o}_{ij}$. In the semantic graph $\boldsymbol{G}_{it}$, each node $\boldsymbol{s}_{ik}$ represents an object attribute. The initial node representation $\boldsymbol{s}_{ik}^{(0)}$ is the feature $\boldsymbol{e}_{ik}$ from visual description module.

Aggregation scheme. After constructing multimodel graph and initializing the representation of each node, we propose two aggregators which guide the information flow between the visual graph and the semantic graph. This aggregation scheme leverages diverse types of contexts from different modalities to refine the node representations, as shown in Figure 2. The first aggregator utilizes attribute features to update the visual nodes. For each node $\boldsymbol{v}_{ij}$ in visual graph $\boldsymbol{G}_{iv}$, the aggregator updates the representation of $\boldsymbol{v}_{ij}$ by attending on neighbour nodes in semantic graph $\boldsymbol{G}_{it}$. Concretely, we first calculate the relevance score between the node $\boldsymbol{v}_{ij}$ and its neighboring node $\boldsymbol{s}_{ik}$ as below:

$$\boldsymbol{r}'_{s_{ik}, v_{ij}} = f_v(\boldsymbol{v}_{ij}^{(0)})^T (f_s(\boldsymbol{s}_{ik}^{(0)})) \tag{1}$$

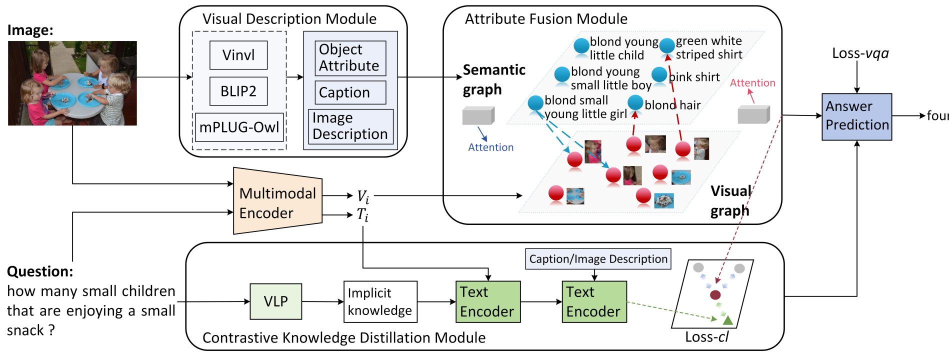


Figure 2: The overview of our attribute-centric approach. Visual description module generates descriptive text for object attributes. Attribute fusion module establishes a multimodal graph and fuses attribute features with visual features by passing messages between two subgraphs. Contrastive knowledge distillation module introduces a series of implicit knowledge to supplement information that cannot be covered in the attributes. On this basis, the contrastive loss is adopted to further strengthen and enrich the representation of attribute features. The blue or red arrows between nodes in the two graphs represent the direction of information flow.

$$\boldsymbol{r}^v_{s_{ik},v_{ij}} = \frac{\exp(\boldsymbol{r}'_{s_{ik},v_{ij}})}{\sum_{s_{ik}\in\mathcal{N}_{v^s_{ij}}} \exp(\boldsymbol{r}'_{s_{ik},v_{ij}})} \tag{2}$$

where f_s and f_v are the multi-layer perceptron (MLPs) used to encode node features. $\mathcal{N}_{v^s_{ij}}$ indicates the neighboring nodes of v_{ij} in the semantic graph. After that, we aggregate the information of attended nodes from the semantic graph to the visual graph. Each visual node is updated:

$$\boldsymbol{v}^{(1)}_{ij} = [\boldsymbol{v}^{(0)}_{ij}; \sum_{s_{ik}\in\mathcal{N}_{v^s_{ij}}} \boldsymbol{r}^v_{s_{ik},v_{ij}} f_{s'}(\boldsymbol{s}^{(0)}_{ik})] \tag{3}$$

where $f_{s'}$ is an MLP that encodes the features of neighboring nodes from the semantic graph. $[;]$ denotes the concatenation of two vectors. Similar to the update mechanism for visual nodes, we further obtain the updated attribute representations as follows:

$$\boldsymbol{r}^s_{s_{ik},v_{ij}} = \frac{\exp(\boldsymbol{r}'_{s_{ik},v_{ij}})}{\sum_{v_{ij}\in\mathcal{N}_{s^v_{ik}}} \exp(\boldsymbol{r}'_{s_{ik},v_{ij}})} \tag{4}$$

$$\boldsymbol{s}^{(1)}_{ik} = [\boldsymbol{s}^{(0)}_{ik}; \sum_{v_{ij}\in\mathcal{N}_{s^v_{ik}}} \boldsymbol{r}^s_{s_{ik},v_{ij}} f_{v'}(\boldsymbol{v}^{(1)}_{ij})] \tag{5}$$

where $f_{v'}$ is an MLP used to encode the features of neighboring nodes. $\mathcal{N}_{s^v_{ik}}$ represents the neighboring nodes of s_{ik} in the visual graph.

Contrastive Knowledge Distillation Module

Contrastive knowledge distillation module aims to further consolidate the representation learning of attribute features. Firstly, we introduce a series of implicit knowledge, and then distill knowledge into attributes through contrastive loss. This further enhances the understanding of scenes, greatly boosting OOD robustness.

Text encoding. Specifically, inspired by prompting GPT-3, we first utilize prompts to acquire implicit knowledge stored in VLP models OFA, BLIP and BLIP2. We adopt the question $\boldsymbol{Q}_i$ and the image $\boldsymbol{I}_i$ as prompt to generate exploratory answers and obtain its word embeddings $\boldsymbol{P}_i$.

To better encourage the alignment between tokens, following compound token attention (Aladago and Piergiovanni 2022), we adopt an enhanced transformer method based on channel fusion to encode features. It maps the question features $\boldsymbol{T}_i$ and implicit knowledge features $\boldsymbol{P}_i$ separately into half of the embedding space:

$$\boldsymbol{P}'_i = f_1(\boldsymbol{P}_i) \tag{6}$$

$$\boldsymbol{T}'_i = f_2(\boldsymbol{T}_i) \tag{7}$$

where f_1 and f_2 are MLPs. Subsequently, we employ two cross-attention layers to independently encode the features and then merge the original features:

$$\hat{\boldsymbol{T}}_i = [\boldsymbol{T}'_i; H_1(\boldsymbol{T}'_i, \boldsymbol{P}'_i, \boldsymbol{P}'_i)] \tag{8}$$

$$\hat{\boldsymbol{P}}_i = [\boldsymbol{P}'_i; H_2(\boldsymbol{P}'_i, \boldsymbol{T}'_i, \boldsymbol{T}'_i)] \tag{9}$$

where $H_1(q,k,v)$, $H_2(q,k,v)$ are two cross-attention function with q, k, and v as query, key and value respectively. Then, it combines the features of the two input streams to create compound tokens and learns the final representation through a self-attention function $G_{att}(x)$:

$$\boldsymbol{Z}_i = G_{att}([\hat{\boldsymbol{T}}_i; \hat{\boldsymbol{P}}_i]) \tag{10}$$

As a result, we obtain the fused features $\boldsymbol{Z}_i$ of the question feature $\boldsymbol{T}_i$ and implicit knowledge $\boldsymbol{P}_i$. This encoding process can adequately focus on question-related knowledge from implicit knowledge. Next, we further fuse the obtained $\boldsymbol{Z}_i$ with image caption $\boldsymbol{C}_i$ to acquire the representation of image-related parts. Consequently, we acquire the encoded knowledge feature $\boldsymbol{F}_i$.

Dataset	#QA pairs	#Images	Image Source
COCO-QA	118K	123K	COCO
TDIUC	1.6M	167K	COCO + VG
VQA-CPv1	370K	205K	COCO
VQA-CPv2	603K	219K	COCO
VQAv2	1.1M	204K	COCO
VQAvs	658K	877K	COCO

Table 1: Comparison of datasets used in this paper. VG represents Visual Genome dataset.

Finally, we utilize two top-down attention networks (Anderson et al. 2018) to obtain question-oriented attribute features and knowledge features, formulated as,

$$\overline{\boldsymbol{S}}_i = f_{att}^s(\boldsymbol{S}_i, \boldsymbol{Z}_i)^T \boldsymbol{S}_i \tag{11}$$

$$\overline{\boldsymbol{F}}_i = f_{att}^t(\boldsymbol{F}_i, \boldsymbol{Z}_i)^T \boldsymbol{F}_i \tag{12}$$

where f_{att}^s and f_{att}^t are top-down attention networks, $\boldsymbol{Z}_i$ is the question feature after transformer encoding.

Contrastive loss. Inspired by the LRC-BERT method (Fu et al. 2021), which employs contrastive learning for latent semantic distillation in the intermediate layers, we use contrastive loss to distill knowledge into attributes. Given question-related attribute features $\overline{\boldsymbol{S}}_i$ and knowledge features $\overline{\boldsymbol{F}}_i$, we construct positive sample pairs $(\overline{\boldsymbol{S}}_i, \overline{\boldsymbol{F}}_i^+)$ and negative sample pairs $(\overline{\boldsymbol{S}}_b, \overline{\boldsymbol{F}}_b)_{b=1}^B$ in the same batch. $(b \neq i)$. B is the number of negative samples in a batch. Following MMBS (Si et al. 2022a), we adopt the cosine similarity as the scoring function. The contrastive loss is formulated as:

$$L_{cl} = -\log \frac{e^{cos(\overline{\boldsymbol{S}}_i, \overline{\boldsymbol{F}}_i^+)}}{e^{cos(\overline{\boldsymbol{S}}_i, \overline{\boldsymbol{F}}_i^+)} + \sum_{b=1}^{B} e^{cos(\overline{\boldsymbol{S}}_b, \overline{\boldsymbol{F}}_b)}} \tag{13}$$

Answer Prediction Module

The answer prediction module takes the question-oriented attribute features $\overline{\boldsymbol{S}}_i$ and knowledge features $\overline{\boldsymbol{F}}_i$ as inputs, and outputs the answer, as follows:

$$\boldsymbol{Y}_i^{pre} = f_{ans}([\overline{\boldsymbol{S}}_i; \overline{\boldsymbol{F}}_i]) \tag{14}$$

where f_{ans} represents an MLP used to calculate the scores for all candidate answers. The overall training objective comprises two components: the VQA multi-label classification loss L_{vqa} and the contrastive loss L_{cl}.

Experiments

Dataset and Experimental Settings

Dataset. We assess the performance of our approach on image understanding datasets (COCO-QA, TDIUC) and OOD datasets (VQA-CPv1, VQA-CPv2, VQAv2 and VQAvs), which validates its capability in addressing image-understanding problem and OOD problem respectively. The dataset statistics can be found in Table 1. For the detailed introduction to the datasets, please refer to Related Work.

Methods	All	Objects	Number	Color	Location
SAN (2016)	61.60	65.40	48.60	57.90	54.00
QRU (2016)	62.50	65.06	46.90	60.50	56.99
HieCoAtt (2016)	65.40	68.00	51.00	62.90	58.80
Dual-MFA (2018)	66.49	68.86	51.32	65.89	58.92
CVA (2018)	67.51	69.55	50.76	68.96	59.93
MCAN (2019)	68.08	69.39	54.19	71.52	60.17
ODA (2018b)	69.33	70.48	54.70	74.17	60.90
CoR (2018a)	69.38	70.42	55.83	74.13	60.57
CAM (2022b)	69.68	70.32	55.26	77.10	59.28
ALSA (2022)	69.97	71.59	54.83	72.74	61.78
MCAN+PA (2022)	70.10	71.13	55.97	74.85	62.07
MRA-Net (2022a)	70.27	71.40	56.42	74.69	60.62
OAM-VQA	**75.22**	**75.67**	**68.20**	**80.66**	**63.80**

Table 2: Comparison with the state-of-the-art approaches on the COCO-QA dataset.

Methods	TDIUC	Methods	VQA-CP v1
MCB (2016)	81.86	SAN (2016)	26.88
SAN (2016)	82.00	NMN (2016)	29.64
RAU (2017)	84.26	MCB (2016)	34.39
QTA (2018)	85.03	Counter (2018)	37.67
BAN (2018)	85.50	GVQA (2018)	39.23
DFAF (2019a)	85.55	UpDn (2018)	39.74
BLOCK (2019)	85.96	LXMERT$_\dagger$ (2019)	52.21
CoR (2018a)	86.91	AdvReg (2018)	43.43
MIRTT (2021)	87.50	RUBi (2019b)	50.90
MLI (2019b)	87.60	LMH (2019)	55.73
MRA-Net (2022a)	87.73	CCS+UpDn (2020)	60.95
DCAF (2019)	88.0	AdaVQA+UpDn (2021a)	61.20
MuRel (2019a)	88.20	CL (2020)	61.27
OAM-VQA	**90.62**	**OAM-VQA**	**65.43**

Table 3: Comparison with the state-of-the-art approaches on the TDIUC and VQA-CPv1 datasets.

Experimental settings. Our model is trained by AdamW optimizer with 100 epochs. The self-attention function $G_{att}(x)$ in the module consists of 5 layers of self-attention. In the cross-attention and self-attention layers, the hidden layer dimension is 512, and the number of heads is 8.

Comparisons with State-of-the-Arts

Comparison on image understanding datasets. We compare our method with the state-of-the-art methods on **COCO-QA** dataset in Table 2. Our proposed method consistently outperforms the state-of-the-art MRA-Net with comfortable margin (70.27% $\sim$ 75.22% absolute accuracy improvement). In particular, OAM-VQA improves performance (from 56.42% $\sim$ to 68.20%) on number-questions. MRA-Net (Peng et al. 2022a) explores various visual relationships to improve model performance, while we bring in object attributes that provide more visual semantic information. This directs the model's focus more towards the objects themselves, *thereby enhancing its ability to handle counting-type questions.* In Table 3, we evaluate our model on the **TDIUC** dataset. The results show that our method achieves the highest performance, specifically surpassing MuRel 2.42%. These findings indicate that *our approach utilizes object attributes to enhance the understanding of visual content, thus excelling in solving object-level fine-grained questions.*

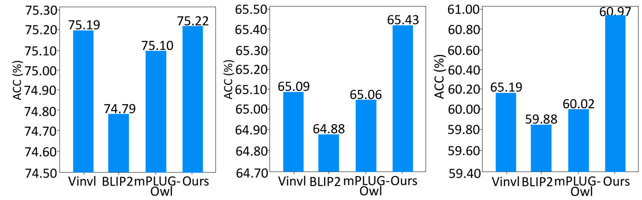
Methods	VQA-CP v2 test				VQAv2 val			
	All	Y/N	Num	Other	All	Y/N	Num	Other
Base models								
SAN (2016)	24.96	38.35	11.14	21.74	52.41	70.06	39.28	47.84
BAN (2018)	37.03	41.55	12.43	41.4	63.9	81.42	45.18	55.54
UpDn (2018)	39.74	42.27	11.93	46.05	63.48	81.18	42.14	55.66
LXMERT†	51.85	54.38	26.92	58.01	70.94	87.92	57.57	61.33
Debiasing methods								
AdvReg (2018)	41.17	65.49	15.48	35.48	62.75	79.84	42.35	55.16
HINT (2019)	46.73	70.04	10.68	46.31	63.38	81.18	42.99	55.56
RUBi (2019b)	47.11	68.65	20.28	43.18	61.16	-	-	-
SCR (2019)	48.47	70.41	10.42	47.29	62.30	77.40	40.90	56.50
LMH (2019)	52.45	69.81	44.46	45.54	61.64	77.85	40.03	55.04
CF-VQA (2021)	53.55	91.15	13.03	44.97	63.54	82.51	43.96	54.3
MMBS* (2022a)	56.51	79.83	28.70	51.92	70.85	88.25	55.67	61.63
CSS (2020)	58.95	84.37	49.42	48.21	59.91	73.25	39.77	55.11
Re-scaling* (2021b)	66.40	79.77	59.06	61.41	69.76	85.32	52.07	62.60
SAR* (2021)	66.73	86.00	62.34	57.84	69.22	87.46	51.20	60.12
MUTANT* (2020)	69.52	93.15	67.17	57.78	70.24	89.01	54.21	59.96
MDDC* (2023b)	69.77	87.88	52.8	64.93	74.51	90.14	58.81	66.76
OAM-VQA	60.97	69.98	53.09	58.41	71.99	88.63	57.94	63.25

Table 4: Comparison with the state-of-the-art approaches on the VQA-CPv2 test and VQAv2 val datasets. † denotes our implementation. $*$ indicates that the models adopt LXMERT as the baseline.

Methods	Language-bias	Visual-bias	Multimodal-bias	Average
S-MRL (2019b)	43.03	31.65	49.48	42.65
UpDn (2018)	47.22	37.35	52.55	46.80
+ LMH (2019)	46.33	37.56	50.75	45.85
+ LMH-L (2019)	47.33	36.08	52.38	46.51
+ LMH-V (2019)	46.68	36.93	52.28	46.38
+ SSL (2021)	45.98	36.43	51.28	45.62
BAN (2018)	48.97	38.51	54.65	48.53
LXMERT†	53.13	41.17	61.05	53.16
OAM-VQA	**53.87**	**42.10**	**61.23**	**53.71**

Table 5: Comparison with the state-of-the-art approaches on the VQAvs dataset. For example, language bias contains keyword bias, visual bias consists of key object bias, and multimodal bias involves combinations of the two.

Comparison on OOD datasets. Table 4 shows the comparison results on the **VQA-CPv2** test. Unlike the datasets mentioned above, the plain VQA models without debiasing methods perform poorly on these biased datasets. Therefore, we compare our model with plain VQA models and debiasing methods. Brief descriptions of baseline models are in Appendix A. For the VQA-CPv2 test, our approach improves the backbone LXMERT with a large performance gain (+9.12%). Specifically, on the number-questions, our model achieves a 26.17% boost. In Table 3, our approach outperforms the CL method by 4.16% on the **VQA-CPv1** dataset. Existing debiasing methods for VQA-CP often rely on its construction characteristic that "the answer distribution under the same question type in the training set and test set are almost reverse" (Si et al. 2022b; Teney et al. 2020). Therefore, the latest SOTA methods like MDDC, SAR and MUTANT can always perform best. In contrast, our method does not use such dataset-specific characteristic and also achieves competitive performance. Besides, most debiasing methods tend to enhance the performance of VQA-CP at the expense of sacrificing the performance of VQAv2 (e.g., CSS, LMH, SAR), while our approach achieves improvements on both datasets, showing genuine out-of-distribution robustness. We also achieve favorable performance on the VQAv2 dataset presented in Table 4, surpassing LXMERT by 1.05%. Table 5 displays the performance for alleviating the language biases, visual biases and multimodal biases on **VQAvs**. In terms of language bias and visual bias, our model outperforms LXMERT by 0.74% and 0.93% respectively. These results demonstrate that *our approach leverages object attributes to enhance the understanding of scenes, thereby boosting the OOD performance.*



(a) Performance on COCO-QA. (b) Performance on VQA-CPv1. (c) Performance on VQA-CPv2.

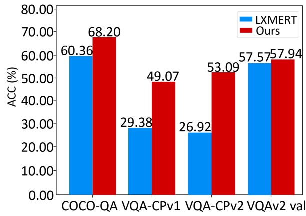
Figure 3: Performance with different types of visual descriptions. Vinvl generates object-level attributes, BLIP2 generates image-level global captions and mPLUG-Owl generates image-level detailed descriptions.

Models	Datasets		
	COCO-QA	VQA-CPv1	VQA-CPv2
LXMERT	72.61	52.21	51.85
LXMERT+CKDM	74.94	63.97	58.33
LXMERT+AFM	75.19	65.09	60.19
LXMERT+AFM+CKDM (ours)	**75.22**	**65.43**	**60.97**

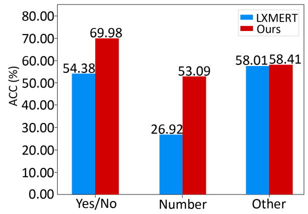
Table 6: Ablation of key components in OAM-VQA on COCO-QA, VQA-CPv1 and VQA-CPv2. "AFM" represents Attribute Fusion Module, and "CKDM" stands for Contrastive Knowledge Distillation Module.

Ablation Study

We conduct ablation studies on the COCO-QA, VQA-CPv1 and VQA-CPv2 datasets to examine the effectiveness of our approach. COCO-QA serves as a representative dataset for image understanding, VQA-CPv1/v2 represent out-of-distribution (OOD) datasets. From Figure 3 and Table 6, several observations can be derived: (1) In the Figure 3, we assess the effectiveness of different levels of descriptive text about visual content. We find that the model with object attribute performs the best. This is because object-level visual-linguistic alignment is more effective than global alignment. In addition, the performance gains brought by image captions are slightly higher than those of image descriptions. (2) In Table 6, we study the ablation of key components of our method. We observe that the attribute fusion module achieves comparative improvements (+2.58% on COCO-QA, +12.88% on VQA-CPv1 and +8.34% on VQA-CPv2) compared to LXMERT. This is because the attribute fusion module effectively fuses object attribute with visual features through a multimodal graph neural network. Besides, we notice that the contrastive knowledge distillation module further enhances the performance. This is because this module introduces a series of textual knowledge to further enrich

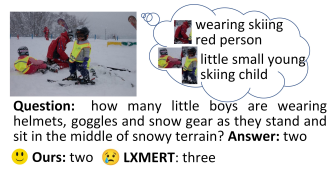


(a) Number-question on four datasets.



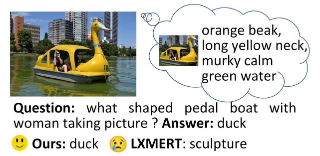
(b) Different types of questions on VQA-CP v2 test.

Figure 4: Performance with different question types. The red bar represents our approach, and the blue bar represents LXMERT.

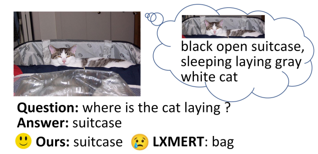


(a) Number-question.

(b) Color-question.



(c) Object-question.



(d) Location-question.

Figure 5: Examples of four different question types on the COCO-QA dataset.

the representation of attribute features and promotes visual-linguistic alignment through contrastive loss. Furthermore, we investigate the impact of different types of knowledge on datasets in Appendix B. We find that implicit knowledge from OFA contributes the most to OOD datasets. The knowledge from BLIP2 has a greater impact on the image understanding datasets. Although both BLIP2 and OFA are visual language pre-training models with encoder-decoder structure, the decoder in BLIP2 is a large language model. *Containing more visual information in the question helps to stimulate more knowledge from the large language model.* Therefore, for the image understanding dataset COCO-QA, BLIP2 offers more efficient knowledge. More detailed examples are shown in Appendix B.

Analysis

Performance on different question types. From Table 2 and Table 4, we investigate the comparison between our approach and LXMERT across different question types, including object, number, color, location, yes/no and other questions. For number questions, our method achieves remarkable improvements of 7.84%, 26.17% and 0.37% on the COCO-QA, VQA-CPv2 and VQAv2 datasets respectively compared to LXMERT. Regarding yes/no questions, our method outperforms LXMERT by +15.60% on VQA-CPv2 and +0.71% on VQAv2 dataset. This also supports our conclusion: *object attribute enhances object-level visual understanding, aiding in addressing object-level fine-grained problem*.

In Figure 4, we further visualize the performance comparison of our approach and LXMERT across different question categories. From Figure 4(a), it is evident that our approach significantly outperforms LXMERT on number-question across all four datasets. In Figure 4(b), for the VQA-CPv2 dataset, our approach outperforms LXMERT by 15.60%, 26.17% and 0.40% on Yes/No, number and other questions respectively. This result demonstrates that *our method excels not only in number-question but also remains highly effective across a broader range of question types and datasets*. Therefore, we conclude that object attribute matters in visual question answering.

Qualitative analysis. In Figure 5, we analyze examples from four question types on the COCO-QA dataset: number, color, object and location. We conclude the following two insights: (1) In multimodal scenarios with noise interference, object attributes enable the model to pay greater attention to question-oriented visual objects. In Figure 5(a), there are some children and an adult. When calculating the number of children, the adult adds complexity to the model. However, our approach uses the attribute fusion module to fuse object attributes and visual features, thereby enhancing the understanding of visual content. Our approach overcomes those interferences and answers the question correctly. In Figure 5(b), the noise is the red rectangular box on the white airplane. The object attributes provide more descriptive information about the airplane, and help our method overcome the noise and understand the overall color of the airplane. However, LXMERT lacks these object-level fine-grained attributes and answers the question incorrectly. (2) For complex scene-understanding questions, object attributes offer valuable answer-related clues. In Figure 5(c) and Figure 5(d), we see that the object attributes provide relevant information about the correct answer, such as orange beak, long yellow neck, and black open suitcase. Our attribute-centric approach effectively fuses these attributes and answers these questions correctly.

Conclusion

In this paper, we propose an effective method to achieve object-level visual-linguistic alignment. Our method designs an attribute fusion module to fuse object attributes with visual features, thus enhancing the understanding of object-level visual content. Subsequently, through the contrastive knowledge distillation module, we introduce a series of implicit knowledge from visual-language pre-trained model, further reinforcing the representation learning of attribute features. Through contrastive loss, we distill knowledge into attributes. This further enhances the understanding of scenes and greatly improves the OOD performance. Extensive experiments conducted on image understanding datasets (COCO-QA and TDIUC) and OOD datasets (VQA-CPv1/v2, VQAv2 val and VQAvs) demonstrate the advantages of our approach. We explore the role of describing visual content text from different levels. We hope that our work will encourage more attention to the understanding of object attribute, promoting the advancement of VQA.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62072212, 61976207), the Development Project of Jilin Province of China (Nos. 20220508125RC, 20230201065GX), and the Jilin Provincial Key Laboratory of Big Data Intelligent Cognition (No. 20210504003GH).

References

Agrawal, A.; Batra, D.; Parikh, D.; and Kembhavi, A. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*.

Aladago, M. M.; and Piergiovanni, A. 2022. Compound Tokens: Channel Fusion for Vision-Language Representation Learning. *arXiv preprint arXiv:2212.01447*.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Andreas, J.; Rohrbach, M.; Darrell, T.; and Klein, D. 2016. Neural module networks. In *CVPR*.

Ben-Younes, H.; Cadene, R.; Thome, N.; and Cord, M. 2019. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *AAAI*.

Cadene, R.; Ben-Younes, H.; Cord, M.; and Thome, N. 2019a. Murel: Multimodal relational reasoning for visual question answering. In *CVPR*.

Cadene, R.; Dancette, C.; Cord, M.; Parikh, D.; et al. 2019b. Rubi: Reducing unimodal biases for visual question answering. In *NeurIPS*.

Chen, L.; Yan, X.; Xiao, J.; Zhang, H.; Pu, S.; and Zhuang, Y. 2020. Counterfactual samples synthesizing for robust visual question answering. In *CVPR*.

Clark, C.; Yatskar, M.; and Zettlemoyer, L. 2019. Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. In *EMNLP-IJCNLP*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.

Fu, H.; Zhou, S.; Yang, Q.; Tang, J.; Liu, G.; Liu, K.; and Li, X. 2021. LRC-BERT: latent-representation contrastive knowledge distillation for natural language understanding. In *AAAI*.

Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In *EMNLP*.

Gao, D.; Li, K.; Wang, R.; Shan, S.; and Chen, X. 2020. Multi-modal graph neural network for joint reasoning on vision and scene text. In *CVPR*.

Gao, F.; Ping, Q.; Thattai, G.; Reganti, A.; Wu, Y. N.; and Natarajan, P. 2022. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In *CVPR*.

Gao, P.; Jiang, Z.; You, H.; Lu, P.; Hoi, S. C.; Wang, X.; and Li, H. 2019a. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *CVPR*.

Gao, P.; You, H.; Zhang, Z.; Wang, X.; and Li, H. 2019b. Multi-modality latent interaction network for visual question answering. In *ICCV*.

Gokhale, T.; Banerjee, P.; Baral, C.; and Yang, Y. 2020. MUTANT: A Training Paradigm for Out-of-Distribution Generalization in Visual Question Answering. In *EMNLP*.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*.

Gui, L.; Wang, B.; Huang, Q.; Hauptmann, A. G.; Bisk, Y.; and Gao, J. 2022. KAT: A Knowledge Augmented Transformer for Vision-and-Language. In *NAACL*.

Guo, Y.; Nie, L.; Cheng, Z.; Ji, F.; Zhang, J.; and Del Bimbo, A. 2021a. AdaVQA: Overcoming Language Priors with Adapted Margin Cosine Loss. In *IJCAI*.

Guo, Y.; Nie, L.; Cheng, Z.; Tian, Q.; and Zhang, M. 2021b. Loss re-scaling VQA: Revisiting the language prior problem from a class-imbalance view. *IEEE Transactions on Image Processing*.

Kafle, K.; and Kanan, C. 2017. An analysis of visual question answering algorithms. In *ICCV*.

Kim, J.-H.; Jun, J.; and Zhang, B.-T. 2018. Bilinear attention networks. In *NeurIPS*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*. PMLR.

Li, L.; Gan, Z.; Cheng, Y.; and Liu, J. 2019. Relation-aware graph attention network for visual question answering. In *ICCV*.

Li, M.; and Moens, M.-F. 2022. Dynamic key-value memory enhanced multi-step graph reasoning for knowledge-based visual question answering. In *AAAI*.

Li, R.; and Jia, J. 2016. Visual question answering with question representation update (QRU). In *NeurIPS*.

Li, Y.; Hu, B.; Zhang, F.; Yu, Y.; Liu, J.; Chen, Y.; and Xu, J. 2023b. A Multi-modal Debiasing Model with Dynamical Constraint for Robust Visual Question Answering. In *Findings of ACL*.

Liang, Z.; Jiang, W.; Hu, H.; and Zhu, J. 2020. Learning to contrast the counterfactual samples for robust visual question answering. In *EMNLP*.

Liu, F.; Liu, J.; Fang, Z.; Hong, R.; and Lu, H. 2019. Densely connected attention flow for visual question answering. In *IJCAI*.

Liu, Y.; Zhang, X.; Zhao, Z.; Zhang, B.; Cheng, L.; and Li, Z. 2022. ALSA: Adversarial Learning of Supervised Attentions for Visual Question Answering. *IEEE transactions on cybernetics*.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical question-image co-attention for visual question answering. In *NeurIPS*.

Lu, P.; Li, H.; Zhang, W.; Wang, J.; and Wang, X. 2018. Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering. In *AAAI*.

Mao, A.; Yang, Z.; Lin, K.; Xuan, J.; and Liu, Y.-J. 2022. Positional attention guided transformer-like architecture for visual question answering. *IEEE Transactions on Multimedia*.

Nguyen, B. X.; Do, T.; Tran, H.; Tjiputra, E.; Tran, Q. D.; and Nguyen, A. 2022. Coarse-to-fine reasoning for visual question answering. In *CVPR*.

Niu, Y.; Tang, K.; Zhang, H.; Lu, Z.; Hua, X.-S.; and Wen, J.-R. 2021. Counterfactual vqa: A cause-effect look at language bias. In *CVPR*.

Peng, L.; Yang, Y.; Wang, Z.; Huang, Z.; and Shen, H. T. 2022a. MRA-Net: Improving VQA Via Multi-Modal Relation Attention Network. *IEEE TPAMI*.

Peng, L.; Yang, Y.; Zhang, X.; Ji, Y.; Lu, H.; and Shen, H. T. 2022b. Answer Again: Improving VQA With Cascaded-Answering Model. *IEEE Transactions on Knowledge and Data Engineering*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Ramakrishnan, S.; Agrawal, A.; and Lee, S. 2018. Overcoming language priors in visual question answering with adversarial regularization. In *NeurIPS*.

Ren, M.; Kiros, R.; and Zemel, R. 2015. Exploring models and data for image question answering. In *NeurIPS*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*.

Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2008. The graph neural network model. *IEEE transactions on neural networks*.

Selvaraju, R. R.; Lee, S.; Shen, Y.; Jin, H.; Ghosh, S.; Heck, L.; Batra, D.; and Parikh, D. 2019. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *ICCV*.

Shi, Y.; Furlanello, T.; Zha, S.; and Anandkumar, A. 2018. Question type guided attention in visual question answering. In *ECCV*.

Si, Q.; Lin, Z.; yu Zheng, M.; Fu, P.; and Wang, W. 2021. Check It Again: Progressive Visual Question Answering via Visual Entailment. In *ACL-IJCNLP*.

Si, Q.; Liu, Y.; Lin, Z.; Fu, P.; and Wang, W. 2023a. Compressing and Debiasing Vision-Language Pre-Trained Models for Visual Question Answering. In *EMNLP*.

Si, Q.; Liu, Y.; Meng, F.; Lin, Z.; Fu, P.; Cao, Y.; Wang, W.; and Zhou, J. 2022a. Towards Robust Visual Question Answering: Making the Most of Biased Samples via Contrastive Learning. In *Findings of EMNLP*.

Si, Q.; Meng, F.; Zheng, M.; Lin, Z.; Liu, Y.; Fu, P.; Cao, Y.; Wang, W.; and Zhou, J. 2022b. Language Prior Is Not the Only Shortcut: A Benchmark for Shortcut Learning in VQA. In *Findings of EMNLP*.

Si, Q.; Mo, Y.; Lin, Z.; Ji, H.; and Wang, W. 2023b. Combo of Thinking and Observing for Outside-Knowledge VQA. In *ACL*.

Song, J.; Zeng, P.; Gao, L.; and Shen, H. T. 2018. From pixels to objects: cubic visual attention for visual question answering. In *IJCAI*.

Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP-IJCNLP*.

Teney, D.; Abbasnejad, E.; Kafle, K.; Shrestha, R.; Kanan, C.; and Van Den Hengel, A. 2020. On the value of out-of-distribution testing: An example of goodhart's law. In *NeurIPS*.

Wang, J.; Ji, Y.; Sun, J.; Yang, Y.; and Sakai, T. 2021. MIRTT: learning multimodal interaction representations from trilinear transformers for visual question answering. In *Findings of EMNLP*.

Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; and Yang, H. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*.

Wu, C.; Liu, J.; Wang, X.; and Dong, X. 2018a. Chain of reasoning for visual question answering. In *NeurIPS*.

Wu, C.; Liu, J.; Wang, X.; and Dong, X. 2018b. Object-difference attention: A simple relational attention for visual question answering. In *Proceedings of the 26th ACM international conference on Multimedia*.

Wu, J.; and Mooney, R. 2019. Self-critical reasoning for robust visual question answering. In *NeurIPS*.

Yang, Z.; He, X.; Gao, J.; Deng, L.; and Smola, A. 2016. Stacked attention networks for image question answering. In *CVPR*.

Ye, Q.; Xu, H.; Xu, G.; Ye, J.; Yan, M.; Zhou, Y.; Wang, J.; Hu, A.; Shi, P.; Shi, Y.; et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Yu, Z.; Yu, J.; Cui, Y.; Tao, D.; and Tian, Q. 2019. Deep modular co-attention networks for visual question answering. In *CVPR*.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*.

Zhang, Y.; Hare, J.; and Prügel-Bennett, A. 2018. Learning to Count Objects in Natural Images for Visual Question Answering. In *ICLR*.

Zhu, X.; Mao, Z.; Liu, C.; Zhang, P.; Wang, B.; and Zhang, Y. 2021. Overcoming language priors with self-supervised learning for visual question answering. In *IJCAI*.

Zhu, Z.; Yu, J.; Wang, Y.; Sun, Y.; Hu, Y.; and Wu, Q. 2020. Mucko: multi-layer cross-modal knowledge reasoning for fact-based visual question answering. In *IJCAI*.