# GRACE: Graph-Based Contextual Debiasing for Fair Visual Question Answering

Yifeng Zhang, Ming Jiang, and Qi Zhao

University of Minnesota, Minneapolis MN 55455, USA
{zhan6987, mjiang}@umn.edu, qzhao@cs.umn.edu

**Abstract.** Large language models (LLMs) exhibit exceptional reasoning capabilities and have played significant roles in knowledge-based visual question-answering (VQA) systems. By conditioning on in-context examples and task-specific prompts, they comprehensively understand input questions and provide answers relevant to the context. However, due to the reliance on in-context examples, LLMs are susceptible to inheriting dataset biases in context descriptions and the provided examples. Innovative methods are required to ensure that LLMs can deliver unbiased yet contextually relevant responses. To tackle this challenge, we present **GRA**ph-based **C**ontextual D**E**biasing (GRACE), a novel graph-based method for debiasing knowledge-based VQA models. This approach consists of two novel and generally applicable components. First, we propose an unsupervised context graph learning method that combats biases by explicitly creating a balanced context graph under the guidance of fairness constraints. Second, building upon the context graph, we consider both semantic features and reasoning processes to enhance prompting with more relevant and diverse in-context examples. Through extensive experimentation on both in-distribution (OK-VQA) and out-of-distribution (VQA-CP, GQA-OOD) datasets, we demonstrate the effectiveness of GRACE in mitigating biases and achieving generalization. Additionally, analyses of the model performance across gender groups demonstrate GRACE's potential impacts on social equity. Our source code is publicly available at `https://github.com/SuperJohnZhang/ContextGraphKVQA`.

**Keywords:** Knowledge VQA, Large Language Model, Debias

## 1 Introduction

Knowledge-based Visual Question Answering (VQA) is an extended VQA task focusing on developing systems to answer questions with knowledge beyond the visual input. Leveraging the extraordinary reasoning capabilities of large language models (LLMs), knowledge-based VQA methods have propelled significant advancements, surpassing human performance on established benchmarks [5, 32]. This achievement is accomplished under a few-shot in-context learning paradigm [37], where models leverage off-the-shelf captioners [10] to generate natural language descriptions of the contextual knowledge [29] obtained

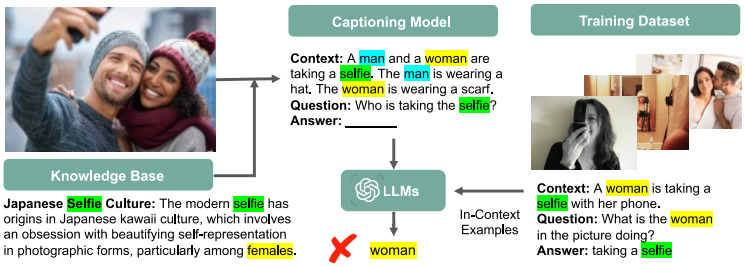



**Fig. 1:** Addressing biases inherited from training data and external knowledge is substantial for visual question answering under the paradigm of in-context learning. Due to potentially biased contexts and in-context examples, LLM reasoners tend to inherit bias and focus on over-represented concepts (*e.g.*, women taking selfies). GRACE mitigates the biases through graph-based context representations and similarity measures.

from the visual input and external sources [22, 38], and then prompt the LLMs with a few in-context examples (*i.e.*, questions and answers with similar contexts) to guide its reasoning [48, 62].

These studies, however, directly generate context descriptions from image features and external knowledge, which are susceptible to biases from the VQA dataset and knowledge bases. For example, as shown in Fig. 1, when posed with the question "Who is taking the selfie?", the knowledge obtained from external sources can be biased (*e.g.*, Japanese selfie culture associates selfies with females). Due to the dominant presence of female-relevant semantics, the captioning model may miss important information about the man (*e.g.*, "the man is holding a phone"). This imbalanced contextual knowledge prevents the captioning model from generating fair contexts and interferes with the search for relevant in-context examples (*e.g.*, more females than males). Consequently, prompts comprising biased context and in-context examples can mislead the LLM to generate incorrect answers.

Bias mitigation in VQA is a longstanding topic. Most VQA debiasing methods focus on addressing dataset biases, typically through balancing data distributions [9] or ensemble learning [12]. These methods cannot effectively address the inherited biases when the model incorporates external knowledge and may not readily adapt to recent LLMs due to the high cost of fine-tuning with counterfactual examples and limited ensemble options [37]. These challenges hinder knowledge-based VQA's generalization and real-world applicability.

To tackle this challenging problem, we introduce **GRA**ph-based **C**ontextual D**E**biasing (GRACE), a novel graph-based approach focusing on providing LLMs with fair context descriptions and in-context examples. It comprises two essential components working together to address biases: First, we mitigate biases in the contexts by constructing a context graph explicitly representing visual features and external knowledge. We introduce novel fairness losses in the learning of the context graph, to ensure a balanced distribution of the graph's structure and the diversity of the incorporated semantic features. Second, we employ the balanced

context graph to find in-context examples for prompting. While existing methods rely on feature similarity in retrieving in-context examples, our context graph enables a holistic evaluation of graph-based semantic similarity and reasoning similarity, considering two distinct dimensions for improving the diversity of in-context examples. Therefore, prompting LLMs with more balanced contexts and diverse in-context examples, GRACE not only fosters fairness within the VQA model but also promotes its generalizability across different test scenarios.

The main contributions of this paper are as follows:

1. Recognizing the bias issue within in-context learning-based VQA models, we take a pioneering step towards mitigating biases through the generation of balanced contexts and the retrieval of diverse in-context examples.
2. We introduce a novel graph-based approach designed to mitigate biases in knowledge-based VQA. It comprises two novel components generally applicable to in-context learning methods: building a balanced context graph and searching for in-context examples based on semantic similarity and reasoning similarity. Together, these components provide a comprehensive solution for addressing biases in knowledge-based VQA.
3. We conduct extensive experiments to assess the effectiveness and applicability of GRACE. It outperforms the state-of-the-art methods across various VQA datasets and base LLMs, including in-distribution (OK-VQA) and out-of-distribution (VQA-CP, GQA-OOD) datasets. Analyses of the performance gap among gender groups highlight the societal benefits of our method.

## 2 Related Work

### 2.1 Visual Question Answering

Existing VQA [5] studies focus on developing datasets and models that cover a broad range of reasoning scenarios, including factual reasoning [5, 20, 32], commonsense reasoning [64], abductive reasoning [26], knowledge-based reasoning [41,58], and reasoning with out-of-distribution data [1,34]. To accommodate the diversity of different settings and enhance the performance of VQA models, advances are made in various aspects, such as cross-modal attention optimization [3, 16, 31, 44, 47, 61], structured inference [4, 11, 23, 27, 30, 31, 49, 63, 67, 68], large-scale pretraining [17,51,52,59]. Recent studies [22,29,38,48] utilize LLMs (*e.g.*, GPT-3 [6]) as reasoners, achieving enhanced reasoning performance. These models operate under the in-context learning paradigm, where visual-linguistic features are described in natural language, namely the context. In-context examples that share a similar context with the test example are retrieved from a training dataset to guide LLMs in predicting answers. For example, PICa [62] employs a state-of-the-art captioning model [66] to convert visual evidence directly into natural language prompts. PromptCap [29] focuses on capturing the correlation between visual details and prompts, enriching the context with better semantics. Prophet [48] uses a vanilla VQA model to generate additional

answer candidates, augmenting the context descriptions. GRACE aligns with recent knowledge-based VQA frameworks [29,48,62] that harness LLM's in-context learning capabilities. In contrast, we focus on addressing the bias problem within the in-context learning paradigm, specifically targeting fairness and generalization for out-of-distribution scenarios. Our graph-based approach mitigates biases in context representation and in-context example retrieval, thereby ensuring the generalizability of knowledge-based VQA models.

### 2.2 Fairness and Bias Mitigation

Fairness in computational systems pertains to assessing equitable treatment for sensitive groups (*e.g.*, gender and race [8]). Mitigating biases is one of the key approaches to fairness. Current bias mitigation methods can broadly fall into three categories: training unbiased models with balanced data [18, 25, 28, 43, 57], unbiased representation learning [40, 69], and specialized training strategies [2, 9, 14, 24, 46, 54, 65]. Previous VQA debiasing studies mainly focus on addressing the biases caused by the imbalanced distribution of training data [1, 9, 34, 69]. These approaches involve using counterfactual examples [9] or ensemble learning [12] for improved generalizability and robustness. While effective on established datasets like VQA-CP [1] and GQA-OOD [34], they cannot systematically address biases in knowledge-based VQA, especially under the in-context learning paradigm. GRACE stands out from existing VQA debiasing methods by addressing biases in both datasets and external knowledge. It achieves this by balancing a context graph with fairness losses and integrating graph-based semantic and reasoning similarity measures to retrieve diverse and relevant in-context examples. This comprehensive approach distinguishes GRACE as a solution for mitigating biases and promoting fairness in reasoning processes within VQA systems.

## 3 Methodology

As a graph-based approach, GRACE facilitates unbiased VQA based on in-context learning and LLMs. It comprises two stages (see Fig. 2):

The first stage aims to generate a balanced context graph representing the visual input $\boldsymbol{x}$ and external knowledge $\boldsymbol{k}$. The context graph is generated with an unsupervised variational autoencoder (see Sec. 3.1) and balanced through the regularization with novel loss terms: semantic fairness loss and structural fairness loss (see Sec. 3.1). It is then translated into natural language descriptions using an off-the-shelf graph-to-caption model [10].

In the second stage, we use the balanced context graph to retrieve in-context examples from the training dataset based on their semantic similarity and reasoning similarity with the test example (see Sec. 3.2). The in-context examples are integrated with the test example into a well-structured prompt. Specifically, each prompt begins with a header explicitly framing the VQA task: "*Please answer the question according to the following examples.*" For in-context examples, the prompt follows a specific template structure, such as: "*Context: [context]*

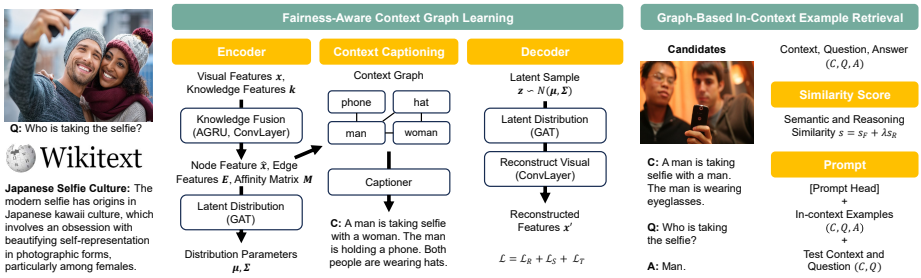



**Fig. 2:** Under the in-context learning paradigm, GRACE addresses biases following a two-stage graph-based approach. The first stage involves generating a balanced context graph using fairness-aware context graph learning, which considers fairness principles regarding graph semantics and structure. In the second stage, we conduct graph-based in-context example retrieval based on semantic similarity and reasoning similarity, and convert the examples into prompts for instructing the LLM reasoner.

*Question: [question] Answer: [answer].*" As for the test example, the template is: "*Context: [context] Question: [question] Answer:*", leaving space for the LLM to generate the answer.

In this section, we introduce our key technical novelties, including fairness-aware context graph learning and graph-based in-context example retrieval. For further details of our approach, please refer to the Supplementary Materials.

### 3.1 Fairness-Aware Context Graph Learning

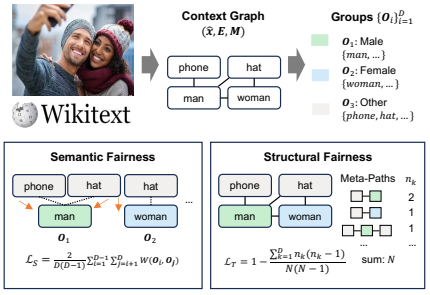
Biases can originate from various sources, including VQA datasets and external knowledge bases. Instead of using conventional debiasing methods such as counterfactual examples [9] or ensemble learning [12], GRACE offers a unique graph-based solution: we explicitly integrate semantic features from the visual input and external knowledge into a context graph, and learn a balanced graph representation with semantic and structural fairness losses.

**Context Graph Generation.** We develop a variational autoencoder network to incorporate external knowledge and generate the context graph in an unsupervised manner. This differs from scene-graph generation methods [21], as it does not require ground-truth graph annotations.

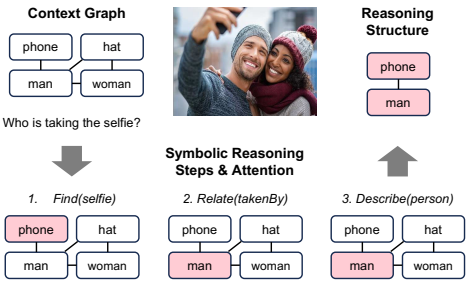
The **encoder** aims to seamlessly integrate the visual features $\boldsymbol{x}$ with external knowledge $\boldsymbol{k}$. We employ an attention-gated recurrent unit (AGRU) [60] network to obtain the knowledge-enriched features

$$\hat{\boldsymbol{x}} = f_{\text{AGRU}}(\boldsymbol{x}, \boldsymbol{k}). \tag{1}$$

While the elements in $\hat{\boldsymbol{x}}$ are considered graph nodes, the edges of the context graph are determined by evaluating the enriched features. For any pair of node




**(a)** Our fairness-aware context graph learning takes into account the semantic fairness ($\mathcal{L}_S$) and structural fairness ($\mathcal{L}_T$) to generate a balanced context graph. Orange arrowed lines denote the information propagation direction.




**(b)** The reasoning process is obtained by aggregating node attention (*i.e.*, the red nodes) across a sequence of reasoning steps generated using a symbolic reasoning model [67].

**Fig. 3:** Illustration of fairness loss and reasoning similarity.

features $\hat{\boldsymbol{x}}_i$ and $\hat{\boldsymbol{x}}_j$, the corresponding edge features $\boldsymbol{e}_{ij} \in \boldsymbol{E}$ and connectivity $m_{ij} \in \boldsymbol{M}$ are computed as

$$\boldsymbol{e}_{ij} = \tanh(f_e(\hat{\boldsymbol{x}}_i \,\|\, \hat{\boldsymbol{x}}_j)), \tag{2}$$

$$m_{ij} = \sigma(f_m(\hat{\boldsymbol{x}}_i \,\|\, \hat{\boldsymbol{x}}_j)), \tag{3}$$

where $f_e(\cdot)$ and $f_m(\cdot)$ are two-layer convolution networks and $\|$ indicates concatenation. Thus, the context graph is represented as the triplet $(\hat{\boldsymbol{x}}, \boldsymbol{E}, \boldsymbol{M})$.

To further encode the context graph into latent distribution, a graph attention network [55] $f_{\text{GAT}}(\cdot)$ and a convolutional network [45] $f_{\text{enc}}(\cdot)$ are employed:

$$\boldsymbol{\mu}, \boldsymbol{\Sigma} = f_{\text{enc}}(f_{\text{GAT}}(\hat{\boldsymbol{x}}, \boldsymbol{M})). \tag{4}$$

Finally, the **decoder** takes samples $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (employing reparameterization trick [36]) and outputs the reconstructed visual features

$$\boldsymbol{x'} = f_{\text{dec}}(\boldsymbol{z}), \tag{5}$$

where $f_{\text{dec}}(\cdot)$ is a convolutional network [21].

**Fairness Objectives.** The above graph generation network is trained end-to-end with the following loss function:

$$\mathcal{L} = \mathcal{L}_S + \mathcal{L}_T + \mathcal{L}_R. \tag{6}$$

In addition to the standard reconstruction loss $\mathcal{L}_R$ that maximizes the evidence lower bound (ELBO) [45], we introduce two regularization loss terms to achieve semantic and structural fairness. The semantic fairness loss $\mathcal{L}_S$ enhances diversity in the graph's semantics, and the structural fairness loss $\mathcal{L}_T$ ensures

balanced graph connectivity. The driving force behind these fairness losses is to untangle superficial correlations between particular concepts and sensitive groups (*e.g.*, the dominant correlations between women and phones in the training images or external knowledge bases, as illustrated in Fig. 3a). The definition of sensitive groups adheres to well-established fairness metrics [7].

The **semantic fairness loss** serves as a regularization factor to enhance semantic fairness within the context graph. It addresses the need for balanced representations across different sensitive groups [13]. For example, it is important to ensure that there is a fair feature distribution between males and females. As illustrated in Fig. 3a, we associate node features $\hat{\boldsymbol{x}}$ and edge features $\boldsymbol{E}$ with $D$ sensitive groups based on their cosine similarity with group labels (*e.g.*, male or female). Representing $\{\boldsymbol{O}_d\}_{d=1}^{D}$ as the $D$ groups of features, $\mathcal{L}_S$ is computed as the average Wasserstein distance $W(\cdot,\cdot)$ between pair combinations of all groups [19], an established fairness metric that measures the "mapping cost" among different embeddings [8, 13].

$$\mathcal{L}_S = \frac{2}{D(D-1)} \sum_{i=1}^{D-1} \sum_{j=i+1}^{D} W(\boldsymbol{O}_i, \boldsymbol{O}_j), \tag{7}$$

Thus, the semantic fairness loss ensures the fair representation of sensitive groups in the context graph.

The **structural fairness loss** enhances the graph fairness by evaluating the diversity of meta-paths [15] within the context graph. A meta-path is an ordered list of the sensitive group labels (*e.g.*, Male, Female, Other) based on the corresponding path from one node to another. For instance, in Fig. 3a, the two relationship paths `phone-man` and `hat-man` are associated with the same meta-path `Other-Male`. Thus, we count the number of paths associated with each meta-path as $n_k$ $(k = 1, 2, \cdots, K)$, where $K$ is the number of meta-paths. The structural fairness loss is computed based on Simpson's Index of Diversity [50]:

$$\mathcal{L}_T = -(1 - \frac{\sum_{k=1}^{D} n_k(n_k - 1)}{N(N-1)}), \tag{8}$$

where $N = \sum_{k-1}^{D} n_k$ is the total number of paths. This structural fairness loss contributes to enhancing the diversity of contexts in the generated graph from the perspective of the graph topology.

### 3.2 Graph-Based In-Context Example Retrieval

Choosing diverse in-context examples is substantial for generating unbiased answers with LLMs. Prior studies retrieve examples from the training dataset based on their similarity with the test example, such as the cosine similarity of image features [62] or multimodal embeddings [29], or the answer similarity in a latent space where a vanilla VQA mode is used to generate answer candidates [48].

These methods search for in-context examples only based on feature similarities, which may result in highly homogeneous examples. Differently, to improve the diversity of in-context examples, we not only retrieve examples using graph-based semantic similarity but also consider the similarity of reasoning processes during question answering.

With the nodes encoding the various semantic concepts in the context graph, the graph-based semantic similarity enables a holistic evaluation of the semantic similarity of contexts, which is distinct from previous similarity measures used. Specifically, given knowledge-enriched node features $\hat{\boldsymbol{x}}^s$, $\hat{\boldsymbol{x}}^t$ of two context graphs, the semantic similarity is computed by measuring the average cosine similarity of each pair of graph nodes

$$s_F(\hat{\boldsymbol{x}}^s, \hat{\boldsymbol{x}}^t) = \frac{1}{|\hat{\boldsymbol{x}}^s||\hat{\boldsymbol{x}}^t|} \sum_{\hat{\boldsymbol{x}}_i^s \in \hat{\boldsymbol{x}^s}} \sum_{\hat{\boldsymbol{x}}_j^t \in \hat{\boldsymbol{x}}^t} \cos(\hat{\boldsymbol{x}}_i^s, \hat{\boldsymbol{x}}_j^t). \tag{9}$$

To measure the reasoning similarity, we employ a symbolic reasoning network [67] that executes a sequence of reasoning functions to arrive at the answer. As shown in Fig. 3b, the reasoning inference is based on the input question (*e.g.*, "who is taking the selfie?" and the constructed context graph. Each reasoning step (*e.g.*, `Find(selfie)`, `Relate(takenBy)`, `Describe(person)`) processes the output of the previous step and computes attention weights over the graph. As a result, the reasoning produces a sequence of intermediate attention weights prioritizing relevant nodes at each step. We aggregate the attention weights across reasoning steps, and prune the context graph by removing reasoning-irrelevant nodes with attention weights below a threshold $\epsilon$. This leads to a compact subgraph (*e.g.*, `shirt-woman-phone`) representing the reasoning process given the input image and question. Based on this structure, given two examples, their reasoning similarity is measured with SimRank [33]:

$$s_R(\mathcal{V}^s, \mathcal{V}^t) = \frac{1}{|\mathcal{V}^s||\mathcal{V}^t|} \sum_{v_i^s \in \mathcal{V}^s} \sum_{v_j^t \in \mathcal{V}^s} \text{SimRank}(v_i^s, v_j^t), \tag{10}$$

where $\mathcal{V}^s$ and $\mathcal{V}^t$ indicate the nodes of the source and target subgraphs, respectively. To holistically measure the semantic and reasoning similarities, we linearly combine them through an empirically assigned weight $\lambda$:

$$s = s_F(\hat{\boldsymbol{x}}^s, \hat{\boldsymbol{x}}^t) + \lambda s_R(\mathcal{V}^s, \mathcal{V}^t). \tag{11}$$

Thus, the combined score considers both the semantic features encoded in the context graph and the reasoning process representing how attention is distributed across the graph, leading to more diverse in-context examples.

## 4 Experiments

In this section, we conduct extensive experiments to demonstrate the effectiveness of GRACE in mitigating biases. For more results and implementation details, we encourage interested readers to explore the Supplementary Materials.

### 4.1 Experimental Settings

**Datasets.** We evaluate models across various VQA benchmarks, considering both in-distribution and out-of-distribution scenarios. In the in-distribution context, we use OK-VQA [41], which emphasizes knowledge-intensive questions. In the out-of-distribution context, we employ VQA-CP [1] and GQA-OOD [34], testing the models' generalization capabilities with unseen data. VQA-CP assesses performance beyond the training data, while GQA-OOD includes distinct head and tail sets for reasoning in unfamiliar scenarios. The comprehensive evaluation using these datasets demonstrates GRACE's performance and adaptability across diverse VQA scenarios. All reported results are from the test sets.

**Compared Models.** For fair comparisons, we rigorously evaluate GRACE by comparing it with state-of-the-art VQA models using GPT-3 as a reasoner (*i.e.*, Prophet [48], PromptCap [29], PICa [62]) or a knowledge source (*i.e.*, REVIVE [38] and KAT [22]). We also evaluate various VQA debiasing methods, including LMH [12] and LMH + CSS [9]. Since LMH and CSS are not directly applicable to LLMs, they are only applied to the captioning model used to generate contexts. Please refer to the Supplementary Materials for details.

**Training.** To train the context graph generation model, we follow REVIVE [38] to incorporate implicit knowledge from GPT-3 [6]) and explicit knowledge from WikiText-2 [42]). With these inputs, we learn the context graph generation with the proposed fairness losses. The training is optimized with an Adam optimizer [35] with a learning rate of 0.0002 and decay rates of $(0.9, 0.999)$.

**Evaluation.** ve We conduct a comprehensive evaluation of VQA models with two major categories of metrics. The bias-related metrics, which are evaluated on standard OOD benchmarks (e.g. GQA-OOD, VQA-CP) with the accuracy of different subsets, quantitatively measure the sources of unfair outcomes. The fairness-related metrics, including word embedding association test (WEAT) [7], VQA-CP gender analyses, and demographic parity analyses [56] (Supp), targeting any prejudice or favoritism toward an individual or group. Both types of evaluation results are highly consistent and provide insights into the models' fairness and reasoning abilities across diverse data distributions.

**Inference Time.** The inference time cost of our method for each sample is comparable to state-of-the-art methods (GRACE: 85ms; Prophet [48]: 76 ms; PromptCap [29]: 74 ms; PICa [62]: 63 ms), where the graph model (27 ms) and symbolic reasoning (18ms) take 53% of the total time cost.

**Hyperparameters.** We use 300-dimensional embeddings to represent nodes and edges of the context graph. The threshold $\epsilon$ is set to 0.3. The number of in-context examples is 20. To balance the importance of semantic similarity and reasoning similarity, we set the weight $\lambda = 2.33$. These hyperparameters are empirically chosen based on validation performance. Please refer to the Supplementary Materials for hyperparameter analyses.

### 4.2 Quantitative Results

Tab. 1 presents quantitative results of various knowledge-based VQA models. Prophet and PromptCap rely on the data and knowledge bias to achieve high

**Table 1:** Comparison with state-of-the-art knowledge-based VQA methods.

| Method | OK-VQA | | VQA-CP | | GQA-OOD | | |
|---|---|---|---|---|---|---|---|
| | WEAT↓ | Acc.↑ | WEAT↓ | Acc.↑ | WEAT↓ | Acc.-T↑ | $\Delta$↓ |
| PICa [62] | 1.59 | 48.06 | 1.60 | 50.08 | 1.74 | 46.82 | 15.57 |
| KAT [22] | 1.64 | 54.41 | 1.67 | 51.94 | 1.69 | 47.98 | 14.32 |
| REVIVE [38] | 1.60 | 58.03 | 1.65 | 52.59 | 1.70 | 48.25 | 17.47 |
| PromptCap [29] | 1.64 | 60.47 | 1.63 | 53.74 | 1.72 | 49.73 | 8.81 |
| Prophet [48] | 1.62 | **61.11** | 1.67 | 53.41 | 1.78 | 49.54 | 9.52 |
| GRACE | **1.52** | 60.32 | **1.51** | **57.35** | **1.63** | **50.14** | **7.49** |

accuracy on OK-VQA, but cannot handle the distribution shift of the VQA-CP and GQA-OOD datasets, resulting in relatively low WEAT scores and test-set accuracy. GRACE, on the other hand, achieves significantly better WEAT scores than all state-of-the-art methods, across all in- and out-of-distribution datasets (*i.e.*, 1.52 on OK-VQA, 1.51 on VQA-CP, and 1.63 on GQA-OOD). It also exhibits the highest accuracy on the out-of-distribution datasets (*i.e.*, 57.35% on VQA-CP, 50.14% on GQA-OOD tail subset, and a 7.49% gap between head and tail). In particular, its performance on VQA-CP is over 3.6% better than state-of-the-art methods. On OK-VQA where the training and test data have similarly biased distributions, GRACE also achieves a competitive 60.32% accuracy. These results demonstrate the capability of the proposed method in handling out-of-distribution test examples where others struggle, highlighting the robustness of our method in finding fair, contextually relevant examples.

### 4.3 Ablation Study

For a comprehensive analysis of GRACE, we conduct ablation studies on how the key components of our method, including the fairness losses and the similarity metrics, contribute to the overall VQA accuracy. For more ablation study results, please refer to the Supplementary Materials.

**Fairness losses.** We apply different combinations of the fairness losses $\mathcal{L}_S$ and $\mathcal{L}_T$ to evaluate their contributions. As shown in Tab. 2, we start with a baseline model that only considers the reconstruction loss $\mathcal{L}_R$ in the learning of context graphs. Adding the semantic fairness loss $\mathcal{L}_S$ or the structural fairness loss $\mathcal{L}_T$, the method achieves promising fairness improvements in terms of WEAT and accuracy metrics, indicating the effectiveness of addressing semantic and structural biases and maintaining accurate graph feature reconstruction. Finally, combining all three losses effectively addresses biases in different aspects and demonstrates significant improvements in the WEAT scores, while maximizing the accuracy on the out-of-distribution VQA-CP and GQA-OOD datasets. This comprehensive approach stands out as the most effective way to promote fairness.

**Reasoning similarity.** GRACE combines semantic and reasoning similarity to enhance in-context example retrieval. On top of the context graph, the ablation study shown in Tab. 2 reveals that selecting in-context examples with

**Table 2:** Ablation study of $\mathcal{L}_S$, $\mathcal{L}_T$, and $s_R$, on top of a strong graph-based baseline model with only the reconstruction loss $\mathcal{L}_R$ and the semantic similarity $s_F$.

| Method | OK-VQA | | VQA-CP | | GQA-OOD | | |
|---|---|---|---|---|---|---|---|
| | WEAT↓ | Acc.↑ | WEAT↓ | Acc.↑ | WEAT↓ | Acc.-T↑ | $\Delta$↓ |
| $\mathcal{L}_R$ $[s_F]$ | 1.65 | 60.36 | 1.62 | 55.38 | 1.73 | 48.45 | 11.09 |
| $+ \mathcal{L}_S$ | 1.57 | 60.14 | 1.53 | 56.87 | 1.66 | 48.92 | 9.98 |
| $+ \mathcal{L}_T$ | 1.54 | 60.22 | 1.54 | 56.48 | 1.69 | 48.56 | 9.51 |
| $+ \mathcal{L}_S + \mathcal{L}_T$ | **1.52** | **60.42** | **1.51** | 56.89 | **1.63** | 48.93 | 9.31 |
| $\mathcal{L}_R$ $[s_F + s_R]$ | 1.65 | 60.31 | 1.62 | 54.87 | 1.73 | 48.85 | 9.94 |
| $+ \mathcal{L}_S$ | 1.57 | 60.25 | 1.53 | 56.68 | 1.66 | 49.28 | 9.27 |
| $+ \mathcal{L}_T$ | 1.54 | 59.94 | 1.54 | 56.74 | 1.69 | 49.05 | 8.73 |
| $+ \mathcal{L}_S + \mathcal{L}_T$ | **1.52** | 60.32 | **1.51** | **57.35** | **1.63** | **50.14** | **7.49** |

**Table 3:** Answer hit rates of the in-context examples. The # of examples is 20.

| Dataset | PICa [62] | PromptCap [29] | Prophet [48] | GRACE | | |
|---|---|---|---|---|---|---|
| | | | | $[s_F]$ | $[s_R]$ | $[s_F + s_R]$ |
| OK-VQA | 56.71 | 78.83 | **82.65** | 79.07 | 68.34 | 81.50 |
| VQA-CP | 51.47 | 64.45 | 63.79 | 74.82 | 60.74 | **79.95** |
| GQA-OOD | 52.57 | 59.46 | 58.79 | 71.57 | 62.64 | **73.65** |

the additional reasoning similarity $s_R$ significantly improves their effectiveness. Especially on GQA-OOD, $s_R$ consistently improves the tail-set accuracy and reduces the accuracy gap $\Delta$, demonstrating its important role in retrieving diverse in-context examples with different semantics but similar reasoning processes, which leads to more effective and unbiased LLM reasoning.

**In-Context Examples.** We further evaluate the effectiveness of in-context example retrieval by computing the hit rate based on the retrieved answers. The hit rate is calculated as the percentage of test samples where the retrieved in-context examples contain the correct answer. Tab. 3 reports the hit rates of state-of-the-art methods and GRACE with different combinations of semantic similarity and reasoning similarity. While state-of-the-art methods cannot achieve high hit rates on VQA-CP and GQA-OOD, all variants of GRACE (*i.e.*, $s_F$, $s_R$, $s_F + s_R$) consistently perform well across all datasets, suggesting the effectiveness of the balanced context graph in mitigating bias. In particular, combining $s_F$ and $s_R$ achieves the highest hit rates, with promising 79.95% on VQA-CP, 73.65% on GQA-OOD, and 81.50% on OK-VQA. These results demonstrate the significant roles of the proposed similarity metrics in retrieving diverse and effective in-context examples.

**LLMs and VLMs.** GRACE is generally applicable to LLMs and Vision-Language Models (VLMs). To demonstrate the effectiveness of GRACE, we compare our approach with Prophet [48] over different reasoners, including LLMs (*e.g.*, GPT-4 [6], Llama2 [53]) and VLMs (*e.g.*, LLaVA-1.5 [39]) in Tab. 4. The

**Table 4:** Performance comparison on different LLMs and VLMs.

| Reasoner | Prophet [48] OK-VQA Acc.↑ | Prophet [48] VQA-CP Acc.↑ | Prophet [48] GQA-OOD Acc.T↑ | Prophet [48] GQA-OOD $\Delta$ ↓ | GRACE OK-VQA Acc.↑ | GRACE VQA-CP Acc.↑ | GRACE GQA-OOD Acc.T↑ | GRACE GQA-OOD $\Delta$ ↓ |
|---|---|---|---|---|---|---|---|---|
| LLaVa-1.5-7B | 58.41 | 52.76 | 48.35 | 8.92 | 58.72 | 54.08 | 48.96 | 8.14 |
| Llama2-7B | 60.24 | 54.02 | 49.45 | 9.48 | 60.38 | 57.32 | 50.23 | 7.85 |
| GPT-3 | 61.11 | 53.41 | 49.54 | 9.52 | 60.32 | 57.35 | 50.14 | 7.49 |
| GPT-4 | **61.17** | 54.09 | 49.66 | 9.74 | 60.46 | **57.61** | **50.21** | **7.23** |

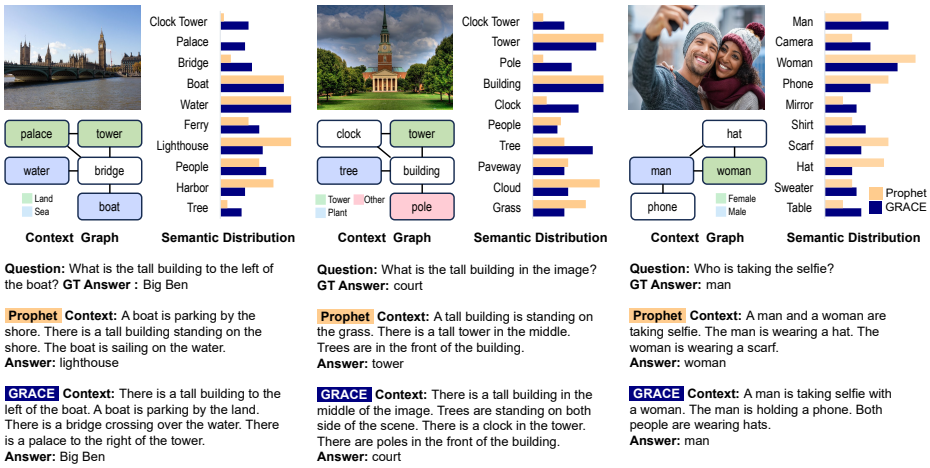



**Fig. 4:** Qualitative comparison between GRACE and Prophet [48]. The node colors in the context graphs denote sensitive groups. The semantic labels with the horizontal bars indicate the number of occurrences in the in-context examples.

results show that GRACE outperforms Prophet on out-of-distribution datasets with all three different LLM/VLM reasoners, achieving the highest performance with GPT-4 on VQA-CP (*i.e.*, 57.61%) and GQA-OOD (*i.e.*, 50.23%). Besides, GRACE archives better performance with 2 out of 3 reasoners on OK-VQA, where the test data maintains a similar distribution as the training data. These results demonstrate the capability of GRACE in generally mitigating biases and conducting fairer reasoning processes.

### 4.4 Qualitative Results

We compare qualitative results between our GRACE and Prophet [48] to evaluate GRACE's bias mitigation and adaptation to out-of-distribution datasets.

Fig. 4 presents the input image, question, ground-truth answer, context graph, and model outputs. It shows that Prophet's context descriptions repetitively mention general knowledge about boats, tall buildings, and the clothing of the man and woman but miss significant details in the images that are relevant to the question-answering. In contrast, our generated context graphs and

**Table 5:** Comparison with VQA debiasing methods.

| Method | OK-VQA | | VQA-CP | | GQA-OOD | | |
|---|---|---|---|---|---|---|---|
| | WEAT↓ | Acc.↑ | WEAT↓ | Acc.↑ | WEAT↓ | Acc.-T↑ | $\Delta$↓ |
| LXMERT [52] | 1.67 | 41.73 | 1.58 | 44.14 | 1.71 | 46.10 | 12.35 |
| + LMH | 1.61 | 38.52 | 1.54 | 52.45 | 1.69 | 45.44 | 8.37 |
| + LMH + CSS | 1.59 | 36.25 | 1.54 | 58.95 | 1.62 | 46.28 | 10.20 |
| Prophet [48] | 1.62 | **61.11** | 1.67 | 53.41 | 1.78 | 49.54 | 9.52 |
| + LMH | 1.58 | 59.96 | 1.64 | 54.33 | 1.74 | 48.36 | 10.41 |
| + LMH + CSS | 1.51 | 57.28 | 1.61 | 55.31 | 1.72 | 48.96 | 9.72 |
| GRACE | 1.52 | 60.32 | 1.51 | 57.35 | 1.63 | 50.14 | **7.49** |
| + LMH | 1.50 | 59.92 | 1.51 | 58.72 | 1.61 | 50.16 | 7.74 |
| + LMH + CSS | **1.44** | 60.15 | **1.44** | **61.37** | **1.61** | **50.26** | 9.25 |

descriptions contain key concepts highly relevant to the answer but are underrepresented in the dataset or knowledge base (*e.g.*, `palace`, `clock`, `pole`, `phone`). This ability to capture answer-relevant details distinguishes GRACE from the previous knowledge-based VQA methods, suggesting its effectiveness in enhancing contextual understanding with fairness-aware context graph learning.

Fig. 4 also visualizes the distribution of semantic concepts in the retrieved in-context examples as bar charts. The semantic concepts are sorted by their cosine similarity to ground-truth answers in descending order. While Phophet retrieves homogeneous in-context examples containing dominant objects in the training dataset (*e.g.*, lighthouse, tower, woman), those retrieved by our approach are not only more diverse (*e.g.*, different types of tall buildings), but also more relevant to the correct answer. These results confirm our approach's ability to balance contexts and retrieve unbiased in-context examples effectively.

### 4.5 Comparison with Debiasing Approaches

To further compare VQA models' capability in mitigating biases, we apply two conventional debiasing methods (*i.e.*, LMH, LMH + CSS) to LXMERT [52], Prophet, and GRACE. As shown in Tab. 5, the fairness of GRACE can be further improved with these methods, achieving the top WEAT score on all datasets. Besides, its out-of-distribution test performances are also improved, with a promising 61.37% accuracy on VQA-CP. In contrast, for LXMERT and Prophet, though the LMH and CSS methods can improve their fairness and out-of-distribution accuracy, their performances on OK-VQA are significantly degraded, suggesting that the high performance of Prophet on OK-VQA is largely attributed to dataset bias. In comparison, GRACE prioritizes the learning and application of balanced context graphs, achieving significantly better performance on VQA-CP and GQA-OOD while maintaining top performance on OK-VQA. Even without additional debiasing techniques, GRACE performs better than Prophet + LMH + CSS across in- and out-of-distribution datasets, suggesting its effectiveness.

**Table 6:** Results on VQA-CP gender subsets.

| Dataset | LXMERT [52] | | | Prophet [48] | | | GRACE | | |
|---|---|---|---|---|---|---|---|---|---|
| | Male↑ | Female↑ | $\Delta$↓ | Male↑ | Female↑ | $\Delta$↓ | Male↑ | Female↑ | $\Delta$↓ |
| Base | 48.65 | 44.28 | 4.37 | 54.80 | 50.41 | 4.39 | 57.62 | 56.72 | 0.90 |
| + LMH | 54.69 | 52.07 | 2.62 | 54.92 | 51.45 | 3.47 | 58.54 | 57.86 | 0.68 |
| + LMH + CSS | **58.41** | **57.28** | **1.13** | **55.47** | **51.62** | **3.85** | **58.92** | **58.31** | **0.61** |

### 4.6 Gender Fairness Assessment

Our work aligns with broader efforts to make AI technologies a positive force for societal well-being. Gender fairness holds particular significance, as it addresses historical and systemic biases that have disproportionately affected individuals based on gender [13]. Following the paradigm of Fair-VQA [43], we cluster the embeddings of questions and answers to create equally sized gender-based subsets of the VQA-CP dataset. We then compare the performance of different models on these subsets. The models are trained on a training set with a higher proportion of male samples (*i.e.*, 63%) and evaluated on a balanced test set. Tab. 6 provides a detailed breakdown of model performances on the male and female subsets. The consistently higher accuracy on the male subset suggests the existence of gender bias (*e.g.*, $\Delta = 4.37\%$ for LXMERT and $\Delta = 4.39\%$ for Prophet). For LXMERT, conventional debiasing methods LMH and CSS significantly improve the VQA accuracy while reducing the between-gender accuracy gap. However, LMH and CSS are less effective on the Prophet model, because they do not contribute directly to the representation of contexts or the selection of in-context examples. GRACE, on the other hand, stands out by reducing the accuracy gap to $\Delta = 0.90\%$. LMH and CSS further improve its overall accuracy, while the gap is minimized to $\Delta = 0.61\%$. These results show our method's potential to effectively mitigate biases that disproportionately affect under-represented groups, indicating its societal significance.

## 5 Conclusion

We have presented GRACE, a novel approach to addressing biases in knowledge-based VQA. It addresses the limitations of existing debiasing methods, which mainly deal with dataset biases and fail to handle biases under the in-context learning paradigm. GRACE features two novel techniques for mitigating biases of VQA models under the in-context learning paradigm: fairness-aware context graph learning and graph-based in-context example retrieval, which aims to create balanced and fair contexts and retrieve diverse in-context examples, allowing for more accurate and unbiased reasoning with LLMs. Experiments on multiple datasets demonstrate the effectiveness of GRACE in handling out-of-distribution scenarios and mitigating biases. Our work advances knowledge-based VQA, offering new avenues for fairness-aware reasoning and in-context learning research.
**Acknowledgment:** This work is supported by NSF Grants 2143197 and 2227450.

## References


1. Agrawal, A., Batra, D., Parikh, D., Kembhavi, A.: Don't just assume; look and answer: Overcoming priors for visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4971–4980 (2018)
2. Alvi, M., Zisserman, A., Nellåker, C.: Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In: Proceedings of the European Conference on Computer Vision Workshops. pp. 0–0 (2018)
3. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6077–6086 (2018)
4. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Neural module networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 39–48 (2016)
5. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
7. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (2017)
8. Caton, S., Haas, C.: Fairness in machine learning: A survey. ACM Computing Surveys (2020)
9. Chen, L., Yan, X., Xiao, J., Zhang, H., Pu, S., Zhuang, Y.: Counterfactual samples synthesizing for robust visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
10. Chen, S., Jin, Q., Wang, P., Wu, Q.: Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9962–9971 (2020)
11. Chen, W., Gan, Z., Li, L., Cheng, Y., Wang, W., Liu, J.: Meta module network for compositional visual reasoning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 655–664 (2021)
12. Clark, C., Yatskar, M., Zettlemoyer, L.: Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. arXiv preprint arXiv:1909.03683 (2019)
13. Dong, Y., Ma, J., Wang, S., Chen, C., Li, J.: Fairness in graph mining: A survey. IEEE Transactions on Knowledge and Data Engineering (2023)
14. Elkan, C.: The foundations of cost-sensitive learning. In: International joint conference on artificial intelligence. vol. 17, pp. 973–978. Lawrence Erlbaum Associates Ltd (2001)
15. Fu, X., Zhang, J., Meng, Z., King, I.: Magnn: Metapath aggregated graph neural network for heterogeneous graph embedding. In: Proceedings of The Web Conference 2020. pp. 2331–2341 (2020)
16. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847 (2016)

17. Gardères, F., Ziaeefard, M., Abeloos, B., Lecue, F.: Conceptbert: Concept-aware representation for visual question answering. In: Proceedings of the Empirical Methods in Natural Language Processing. pp. 489–498 (2020)
18. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Iii, H.D., Crawford, K.: Datasheets for datasets. Communications of the ACM **64**(12), 86–92 (2021)
19. Givens, C.R., Shortt, R.M.: A class of wasserstein metrics for probability distributions. Michigan Mathematical Journal **31**(2), 231–240 (1984)
20. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6904–6913 (2017)
21. Gu, J., Zhao, H., Lin, Z., Li, S., Cai, J., Ling, M.: Scene graph generation with external knowledge and image reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1969–1978 (2019)
22. Gui, L., Wang, B., Huang, Q., Hauptmann, A., Bisk, Y., Gao, J.: Kat: A knowledge augmented transformer for vision-and-language. arXiv preprint arXiv:2112.08614 (2021)
23. Gupta, N., Lin, K., Roth, D., Singh, S., Gardner, M.: Neural module networks for reasoning over text. arXiv preprint arXiv:1912.04971 (2019)
24. Han, X., Wang, S., Su, C., Huang, Q., Tian, Q.: General greedy de-bias learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(8), 9789–9805 (2023)
25. Hendricks, L.A., Burns, K., Saenko, K., Darrell, T., Rohrbach, A.: Women also snowboard: Overcoming bias in captioning models. In: Proceedings of the European conference on computer vision. pp. 771–787 (2018)
26. Hessel, J., Hwang, J.D., Park, J.S., Zellers, R., Bhagavatula, C., Rohrbach, A., Saenko, K., Choi, Y.: The abduction of sherlock holmes: A dataset for visual abductive reasoning. In: European Conference on Computer Vision. pp. 558–575. Springer (2022)
27. Hu, R., Andreas, J., Darrell, T., Saenko, K.: Explainable neural computation via stack neural module networks. In: IEEE trans. pp. 53–69 (2018)
28. Hu, X., Wang, H., Dube, S., Vegesana, A., Yu, K., Lu, Y.H., Yin, M.: Discovering biases in image datasets with the crowd. Proceedings of HCOMP (2019)
29. Hu, Y., Hua, H., Yang, Z., Shi, W., Smith, N.A., Luo, J.: Promptcap: Prompt-guided task-aware image captioning. arXiv preprint arXiv:2211.09699 (2022)
30. Hudson, D., Manning, C.D.: Learning by abstraction: The neural state machine. Advances in Neural Information Processing Systems **32** (2019)
31. Hudson, D.A., Manning, C.D.: Compositional attention networks for machine reasoning. arXiv preprint arXiv:1803.03067 (2018)
32. Hudson, D.A., Manning, C.D.: 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. pp. 6700–6709 (2019)
33. Jeh, G., Widom, J.: Simrank: a measure of structural-context similarity. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 538–543 (2002)
34. Kervadec, C., Antipov, G., Baccouche, M., Wolf, C.: Roses are red, violets are blue... but should vqa expect them to? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2776–2785 (2021)
35. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017)

36. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
37. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. Advances in neural information processing systems **35**, 22199–22213 (2022)
38. Lin, Y., Xie, Y., Chen, D., Xu, Y., Zhu, C., Yuan, L.: Revive: Regional visual representation matters in knowledge-based visual question answering. arXiv preprint arXiv:2206.01201 (2022)
39. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023)
40. Locatello, F., Abbati, G., Rainforth, T., Bauer, S., Schölkopf, B., Bachem, O.: On the fairness of disentangled representations. Advances in neural information processing systems **32** (2019)
41. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3195–3204 (2019)
42. Merity, S., Xiong, C., Bradbury, J., Socher, R.: Pointer sentinel mixture models. arXiv preprint arXiv:1609.07843 (2016)
43. Park, S., Hwang, S., Hong, J., Byun, H.: Fair-vqa: Fairness-aware visual question answering through sensitive attribute prediction. IEEE Access **8**, 215091–215099 (2020)
44. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
45. Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A., Carin, L.: Variational autoencoder for deep learning of images, labels and captions. Advances in neural information processing systems **29** (2016)
46. Ryu, H.J., Adam, H., Mitchell, M.: Inclusivefacenet: Improving face attribute detection with race and gender diversity. arXiv preprint arXiv:1712.00193 (2017)
47. Santoro, A., Raposo, D., Barrett, D.G., Malinowski, M., Pascanu, R., Battaglia, P., Lillicrap, T.: A simple neural network module for relational reasoning. In: Proceedings of the Conference and Workshop on Neural Information Processing Systems. pp. 4967–4976 (2017)
48. Shao, Z., Yu, Z., Wang, M., Yu, J.: Prompting large language models with answer heuristics for knowledge-based visual question answering. In: Computer Vision and Pattern Recognition (CVPR). pp. 14974–14983 (2023)
49. Shi, J., Zhang, H., Li, J.: Explainable and explicit visual reasoning over scene graphs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8376–8384 (2019)
50. Simpson, E.H.: Measurement of diversity. nature **163**(4148), 688–688 (1949)
51. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: Vl-bert: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530 (2019)
52. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490 (2019)
53. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov,

T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open foundation and fine-tuned chat models (2023)
54. Tzeng, E., Hoffman, J., Darrell, T., Saenko, K.: Simultaneous deep transfer across domains and tasks. In: Proceedings of the IEEE international conference on computer vision. pp. 4068–4076 (2015)
55. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)
56. Verma, S., Rubin, J.: Fairness definitions explained. In: Proceedings of the international workshop on software fairness. pp. 1–7 (2018)
57. Wang, A., Liu, A., Zhang, R., Kleiman, A., Kim, L., Zhao, D., Shirai, I., Narayanan, A., Russakovsky, O.: Revise: A tool for measuring and mitigating bias in visual datasets. International Journal of Computer Vision **130**(7), 1790–1810 (2022)
58. Wang, P., Wu, Q., Shen, C., Dick, A., Van Den Hengel, A.: Fvqa: Fact-based visual question answering. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(10), 2413–2427 (2017)
59. Wu, J., Lu, J., Sabharwal, A., Mottaghi, R.: Multi-modal answer validation for knowledge-based vqa. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2712–2721 (2022)
60. Xiong, C., Merity, S., Socher, R.: Dynamic memory networks for visual and textual question answering. In: International conference on machine learning. pp. 2397–2406. PMLR (2016)
61. Xu, H., Saenko, K.: Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In: IEEE trans. pp. 451–466. Springer (2016)
62. Yang, Z., Gan, Z., Wang, J., Hu, X., Lu, Y., Liu, Z., Wang, L.: An empirical study of gpt-3 for few-shot knowledge-based vqa. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 3081–3089 (2022)
63. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21–29 (2016)
64. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: CVPR. pp. 6713–6724 (2019)
65. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. pp. 335–340 (2018)
66. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: Vinvl: Revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5579–5588 (2021)
67. Zhang, Y., Jiang, M., Zhao, Q.: Explicit knowledge incorporation for visual reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1356–1365 (2021)
68. Zhang, Y., Jiang, M., Zhao, Q.: Query and attention augmentation for knowledge-based explainable reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15576–15585 (2022)
69. Zhu, B., Niu, Y., Lee, S., Hur, M., Zhang, H.: Debiased fine-tuning for vision-language models by prompt regularization. arXiv preprint arXiv:2301.12429 (2023)