GRACE: Graph-Based Contextual Debiasing for Fair Visual Question Answering (Supplementary Materials)

Yifeng Zhang, Ming Jiang, and Qi Zhao

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

1 Introduction

In our main paper, we have presented GRACE, a novel approach addressing biases in knowledge-based VQA. It surpasses current debiasing methods, which primarily tackle dataset biases but fall short in handling biases within the in-context learning paradigm. GRACE introduces two novel techniques for mitigating biases in VQA models under in-context learning: fairness-aware context graph learning and graph-based in-context example retrieval. These techniques aim to establish balanced and fair contexts, retrieving diverse in-context examples for more accurate and unbiased reasoning with LLMs. Experimental results across various datasets demonstrate GRACE's effectiveness in handling out-of-distribution scenarios and mitigating biases. The supplementary material provides additional experimental results and implementation details of our proposed work:

1. Sec. 2 provides additional ablation analyses of the model components and hyperparameters used in the in-context example retrieval.
2. Sec. 3 compares the demographic parity of GRACE and the state-of-the-art approaches over different sensitive groups.
3. Sec. 4 presents implementation details of applying debiasing approaches (*i.e.*, LMH [5] and CSS [2]) to the graph-to-caption model [3].
4. Sec. 5 presents a prompt example used to instruct the LLM reasoner.

2 Additional Ablation Study

To optimize the model performance, we conduct an ablation study with different model components and hyperparameters (*i.e.*, values ϵ, λ, and the number of in-context examples). We evaluate model performance on OK-VQA, VQA-CPv2, and GQA-OOD datasets.

Model Components. To assess the generalizability of GRACE, we conducted ablation studies on both GPT-3 and GPT-4 across VQA-CP and GQA-OOD datasets (results in Table 1). Here, we explore the impact of individual components within the model:

The first observation highlights that GRACE exhibits limited sensitivity to the specific choice of symbolic reasoning model (*e.g.*, NSM [6], ProTo [14], MMN [4]) or graph-to-caption model (*e.g.* SGAE [12]). All variants achieved comparable performance on both models (*i.e.*, GPT-3 and GPT-4) and datasets (VQA-CP and GQA-OOD). This suggests that GRACE is relatively agnostic to the underlying implementation details of these modules as long as they effectively capture symbolic reasoning and knowledge representation.

We further analyzed the effect of prompts designed following Prophet's approach [10]. The ablation study reveals two critical insights: First, removing context information from the prompts significantly deteriorates performance (*e.g.*, with GPT-4, the accuracy decreases from 57.61 to 53.68 on VQA-CP, and from 50.21 to 48.09 on GQA-OOD. This drop suggests the necessity of incorporating balanced contextual information within the prompts to guide the LLMs towards effective reasoning. Second, excluding the prompt head from the model also leads to decreased accuracy (*e.g.*, with GPT-4, the accuracy decreases from 57.61 to 56.97 on VQA-CP and from 50.21 to 49.26 on GQA-OOD). This indicates that the prompt head plays a crucial role in translating the knowledge graph and symbolic reasoning steps into well-formulated prompts that the LLMs can understand and leverage for reasoning.

These ablation studies provide valuable insights into the inner workings of GRACE. The model's robustness to variations in symbolic/caption models demonstrates its flexibility, while the critical role of contextual prompts underscores the importance of guiding LLMs with balanced information for effective reasoning.

Table 1: Ablation study of model components on GPT-3 and GPT-4.

Variant	GPT-3			GPT-4		
	VQA-CP	GQA-OOD		VQA-CP	GQA-OOD	
	Acc.	Acc.T	$\Delta\downarrow$	Acc.	Acc.T	$\Delta\downarrow$
Baseline ($\mathcal{L}_R$ [s_F])	55.38	48.45	11.09	55.41	48.42	10.74
w/ NSM [6]	56.98	50.11	8.82	57.48	50.23	8.15
w/ ProTo [14]	57.01	50.04	8.33	57.53	50.16	7.84
w/ MMN [4]	57.04	50.08	7.92	57.42	50.14	7.56
w/ SGAE [12]	57.26	49.95	8.06	57.18	50.02	8.26
w/o context	54.71	47.86	10.14	53.68	48.09	9.93
w/o prompt head	55.98	49.21	9.70	56.97	49.26	8.64
Full	**57.35**	**50.14**	**7.49**	**57.61**	**50.26**	**7.23**

Reasoning Similarity Threshold ϵ. The hyperparameter ϵ is the threshold that determines whether a node of the context graph is relevant for computing the reasoning similarity. On the one hand, a high threshold ϵ results in a more

Table 2: Ablation study of hyperparameter ϵ.

ϵ	OK-VQA	VQA-CP	GQA-OOD	
	Acc. ↑	Acc. ↑	Acc.-T ↑	Δ ↓
0.1	60.04	56.83	49.86	9.35
0.3	**60.32**	**57.35**	**50.14**	**7.49**
0.5	59.74	57.18	50.06	8.83
0.7	56.69	55.48	48.95	10.54
0.9	54.28	53.09	44.75	14.96

Table 3: Ablation study of hyperparameter λ.

λ	OK-VQA	VQA-CP	GQA-OOD	
	Acc. ↑	Acc. ↑	Acc.-T ↑	Δ ↓
9.0	60.17	55.87	48.92	7.74
2.3	**60.32**	**57.35**	**50.14**	**7.49**
1.0	58.73	56.42	47.85	8.84
0.4	56.09	53.89	45.92	10.73
0.1	56.78	54.53	47.02	10.48

compact reasoning process representation, which may exclude important nodes. On the other hand, a low threshold ϵ could result in the inclusion of less relevant nodes. As shown in Tab. 2, the highest accuracy is consistently achieved on all datasets when ϵ is set to 0.3. This observation shows the significance of an appropriately tuned threshold for computing the reasoning similarity.

Similarity Balencing Weight λ. The hyperparameter λ determines the weight of reasoning similarity when combined with the semantic similarity for in-context example retrieval. As shown in Tab. 3, the performance varies with different values of λ. Notably, the model achieves its peak across all evaluation metrics when λ is set to 2.3, suggesting that the reasoning similarity plays a significant role in the retrieval of diverse in-context examples.

Number of In-Context Examples. The number of in-context examples facilitates conditional reasoning with LLMs. The decision on the optimal number of examples involves a balance between performance and efficiency. While increasing this number is anticipated to provide the reasoner with richer knowledge, it concurrently expands the semantic space, imposing a higher demand on the reasoner to navigate and analyze the prompt throughout the reasoning process. The results in Tab. 4 present the empirical exploration of varying numbers of in-context examples. With 20 in-context examples per prompt, our model achieves the highest accuracy across all evaluated metrics. Moreover, a noteworthy reduction in the performance gap Δ is observed when the number of in-context examples exceeds 20. Therefore, we empirically select 20 in-context examples for optimal model performance.

Table 4: Ablation study of the number of examples.

Num.	OK-VQA	VQA-CP	GQA-OOD	
	Acc. ↑	Acc. ↑	Acc.-T ↑	Δ ↓
0	48.52	51.47	41.67	10.38
1	53.39	52.10	43.06	10.57
8	56.61	54.83	47.24	9.80
16	59.48	56.94	49.84	7.85
20	**60.32**	**57.35**	**50.14**	**7.49**
> 20	≥ 59.24	≥ 56.78	≥ 49.74	≥ 8.37

Table 5: Ablation study of fairness losses $\mathcal{L}_S$, $\mathcal{L}_T$, and the reasoning similarity s_R, based on the demographic parity. The baseline context graphs are generated with only the reconstruction loss $\mathcal{L}_R$ and the semantic similarity s_F.

Method	OK-VQA↓	VQA-CP↓	GQA-OOD↓
$\mathcal{L}_R$ $[s_F]$	2.33	2.59	2.64
$+ \mathcal{L}_S$	2.08	2.17	2.53
$+ \mathcal{L}_T$	2.10	2.13	2.51
$+ \mathcal{L}_S + \mathcal{L}_T$	2.11	2.11	2.47
$\mathcal{L}_R$ $[s_F + s_R]$	2.24	2.35	2.56
$+ \mathcal{L}_S$	2.09	2.16	2.42
$+ \mathcal{L}_T$	2.13	2.21	2.47
$+ \mathcal{L}_S + \mathcal{L}_T$	**2.06**	**2.08**	**2.34**

3 Demographic Parity Analysis

In addition to overall fairness metrics detailed in the main paper (*i.e.*, WEAT [1] and accuracy on out-of-distribution datasets), we delve into nuanced investigations of societal bias issues, such as gender/age discrimination. Specifically, we calculate average demographic parity scores among various sensitive groups [9], including gender (*i.e.*, male, female), environmental factors (*i.e.*, indoor, outdoor, *etc.*), cultural aspects (*i.e.*, religion, science, *etc.*), and natural categories (*i.e.*, animal, plant, *etc.*), extending the analysis presented in the main paper. Using the same experimental setup as in the main paper, we compare results across different combinations of technical components and traditional debiasing methods (LMH and CSS), validating the role of GRACE in mitigating bias.

3.1 Contributions of Model Components

Tab. 5 presents the average demographic parity about different combinations of loss terms and similarity measures. In comparison to a baseline with only the reconstruction loss $\mathcal{L}_R$ (first panel), the incorporation of semantic fairness loss $\mathcal{L}_S$ and structural fairness loss $\mathcal{L}_T$ yields noteworthy improvements in mitigating performance differences across distinct sensitive groups. The optimal

Table 6: Demographic parity of different model combinations on OK-VQA, VQA-CP, and GQA-OOD dataset.

Method	OK-VQA↓	VQA-CP↓	GQA-OOD↓
LXMERT [11]	2.19	2.37	2.64
+ LMH	2.14	2.18	2.33
+ LMH + CSS	2.06	2.09	2.28
Prophet [10]	2.34	2.68	2.86
+ LMH	2.30	2.63	2.81
+ LMH + CSS	2.16	2.55	2.69
GRACE	2.06	2.08	2.34
+ LMH	2.01	2.04	2.29
+ LMH + CSS	**1.90**	**1.78**	**2.24**

performance is achieved when both fairness losses are combined, indicating the important role of considering both semantic and structural fairness in the context graph to address biases effectively.

Furthermore, the introduction of reasoning similarity s_R marks important progress across various context graphs. This finding indicates the ability of reasoning similarity to enhance fairness in in-context example retrieval. The joint consideration of semantic and structural fairness, coupled with the inclusion of reasoning similarity, emerges as a comprehensive strategy for improving fairness and mitigating biases in the context of demographic parity.

3.2 Comparison with Debiasing Methods

Tab. 6 reports the average demographic parity with respect to different debiasing approaches (*i.e.*, LMH, CSS) over multiple baselines (*i.e.*, LXMERT, Prophet, GRACE). Notably, our GRACE model, even without debiasing interventions, demonstrates commendable bias mitigation capabilities, achieving the best demographic parity across the evaluated datasets, and showing the intrinsic ability of GRACE to address biases. Furthermore, the subsequent improvements with the application of LMH and CSS highlight their effectiveness in augmenting the model's fairness.

3.3 Gender and Age Parity

Fairness across various sensitive groups holds profound societal implications in real-world scenarios. In this section, we zoom into specific sensitive groups and evaluate the parity among gender and age. To establish a fine-grained analysis of gender and age fairness, we categorize the VQA-CP dataset into different topics (*i.e.*, occupations, sports, activities, others), and evaluate their demographic parity in Tab. 7.

Table 7: Demographic parity of VQA models among gender and age groups.

Method	Gender				
	Avg.	Occupations	Sports	Activities	Others
Prophet [10]	2.95	2.93	1.74	2.56	4.57
PromptCap [8]	2.53	3.14	2.28	1.15	3.55
PICa [13]	2.43	2.87	1.73	0.94	4.18
GRACE	**1.83**	**2.51**	**0.17**	**1.13**	**3.51**
Method	Age				
	Avg.	Occupations	Sports	Activities	Others
Prophet [10]	1.58	1.96	1.08	1.44	1.84
PromptCap [8]	1.61	1.44	0.65	1.81	2.54
PICa [13]	1.74	1.58	0.41	1.32	3.65
GRACE	**0.77**	**0.69**	**0.13**	**0.63**	**1.57**

Analysis of Gender Groups. A long-standing real-world fairness issue is the bias among different gender groups (*i.e.*, male and female). For example, a common stereotype may mistakenly correlate "nurse" to the female group and "doctor" to the male group. The evaluation of demographic parity among gender groups, as depicted in Tab. 7, shows distinct performance levels across various VQA models. Prophet exhibits an average gender demographic parity score of 2.95. PromptCap and PICa have slightly lower gender parity of 2.53 and 2.43, respectively. GRACE demonstrates significantly reduced gender parity with an overall score of 1.83. Particularly noteworthy is GRACE's exceptional performance in specific topics, achieving a minimal score of 0.17 in the "sports" category, highlighting its effectiveness in mitigating gender biases across diverse contexts.

Analysis of Age Groups. Similarly, the evaluation of demographic parity among age groups (*i.e.*, young and old) shows a significant improvement achieved by our GRACE model. Prophet, PromptCap, and PICa exhibit average age demographic parity scores of 1.58, 1.61, and 1.74, respectively. GRACE, however, achieves an outstanding average score of 0.77. Notably, GRACE achieves remarkable parity scores in specific topics, with minimal values of 0.13 in the "sports", 0.63 in the "activities" category, and 0.69 in the "occupations" category, emphasizing its efficacy in mitigating age-related biases across diverse domains.

Case Studies. In specific case studies, GRACE exhibits significant improvements in fairness for certain question categories:

– *Gender and Sports:* Sports-related questions often exhibit gender bias, where stereotypes influence the association of specific sports with a particular gender. In this context, men are predominantly associated with aggressive and intensive games like boxing and American football, while women are often associated with gymnastics and figure skating. The bias is evident in image datasets as well as external knowledge bases. As shown in Fig. 1a, for

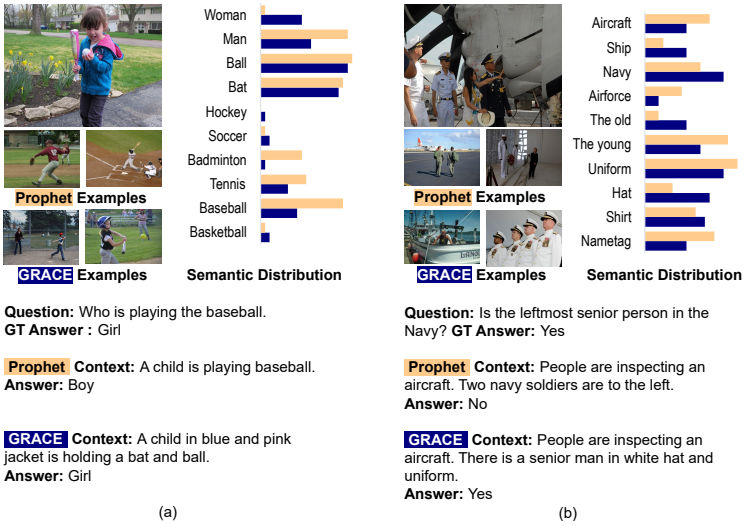

Fig. 1: Supplementary Qualitative Examples for Case Studies. We show the questions, captions, answers, representative in-context examples, and semantic distribution of the Prophet [10] and GRACE.

the question "Who is playing baseball?", existing LLM-based VQA models such as Prophet tend to gather in-context examples primarily about male players, leading to biased responses. This disparity in accuracy between questions about males and females can exacerbate gender stereotypes in sports. GRACE addresses the issue with the search for more diverse in-context examples. Fig. 1(a) shows that by considering the similarity of reasoning processes, specifically the process of identifying certain concepts given visual features of sports equipment, it successfully retrieves more examples of female involvement in baseball, achieving improved gender parity for sports-related questions.

– *Age and Occupations:* Occupational age discrimination poses another practical fairness concern. Some occupations tend to favor specific age groups, exemplified by the preference for young and energetic attributes in roles like Navy soldiers. This bias leads to the under-representation of minority groups, such as older individuals, resulting in an imbalanced distribution of examples across different age groups. In Fig. 1(b), an instance is presented where the question "Is the leftmost senior person in the Navy?" is posed. Different from Prophet, GRACE effectively addresses age biases by considering more visual details, such as the attire of the individual (*e.g.*, hat, uniform), during in-context example retrieval. This approach conditions on more balanced samples, enhancing the fairness of decision-making. Consequently, GRACE exhibits improved fairness in answering occupation-related questions, leading to a substantial reduction in age parity.

4 Debiasing Contexts with LMH and CSS

The experiment in the main paper compares the performance of debiasing models over LLM-based reasoning models. While LMH and CSS are not directly applicable to LLMs, we apply them to the captioner used to generate context descriptions. In this section, we detail how the ensemble heuristic of LMH [5] and the counterfactual examples synthesize mechanism of CSS [2] are adjusted for the off-the-shelf graph-to-caption model [3].

4.1 Applying LMH to Captioner

LMH is an ensemble-based general debiasing approach that mitigates the inherited biases from the training dataset. It consists of two major steps, the training of a bias-only model, and the training of the main model based on the bias-only one. By excluding the shallow correlations modeled in the bias-only model, the main model is expected to be robust across out-of-domain sets.

While VQA and captioning tasks are different by nature, they can share intermediate visual-linguistic features for output generation. Therefore, we train a robust captioner by formulating a multi-task learning paradigm, where a robust VQA answer classifier and robust captioner are simultaneously trained on intermediate features from the input. To train a bias-only model, we follow LMH to take the question type as the input, and train it on the corresponding training set for different datasets (*i.e.*, VQA, GQA, and OK-VQA). Next, by combining the logits of the bias-only model with the robust VQA model with Learned-Mixin [5], we train the robust model on the training set (*i.e.*, VQA, GQA, and OK-VQA). The resulting captioner is expected to exclude the shallow correlations represented in the bias-only VQA model.

4.2 Applying CSS to Captioner

CSS centers around the creation of counterfactual examples to train the model. Given a set of images, questions, and answers, CSS masks either key visual (V-CSS) or linguistic components (Q-CSS) from the input (*i.e.*, image, question) and dynamically updates the output (*i.e.*, answer) accordingly. The new synthesized examples are fed to the original model and mitigate the inherited biases from the dataset.

In this work, to create synthesized samples for the captioner, we adopt V-CSS [2] and extend its dynamic answer update mechanism to work for captions. The idea behind the mechanism is to remove the linguistic description from the ground-truth caption that is correlated to the masked visual regions. Specifically, after selecting a set of critical objects following V-CSS, we first produce their captions by feeding the regional features into the captioner, With the POS Tagger [7] extracting the nouns from those captions, we compute their cosine similarity score with each token in the original ground-truth caption, and replace it with a mask tag "[MASK]" if the score exceeds an empirical threshold ϵ'. We optimize the threshold for best performance.

To debias the captioner for multiple datasets in our experiments (*i.e.*, OK-VQA, VQA-CP, and GQA-OOD), we curate synthesized samples based on the annotations of images from their corresponding training set (*i.e.*, OK-VQA, VQA, GQA). The off-the-shelf captioner is then fine-tuned on the corresponding synthetic samples to mitigate bias for the datasets.

5 Prompt Example

In the following, we provide an example prompt that consists of 8 in-context examples. This prompt corresponds to the first qualitative example in the main paper.

```
Please answer the question according to the following examples. The
provided example may provide relevant contexts for the answer. The answer
should contain one phrase or term.
==
Question: What is the name of the red building?
Context: There is a boat sailing on the sea. A red tower is sitting
on the land.
Answer: lighthouse.
==
Question: What is the bridge crossing over the River Thames?
Context: There is a bridge crossing the river. The Westminster Palace
is by the River Thames.
Answer: Westminster Bridge.
==
Question: What is the name of the river?
Context: A boat is sailing along the shore of the River Thames. Answer:
River Thames
==
Question: What is the purpose of the clock on the tower?
Context: There is a tower located in the church. People are holding
ceremonies in the church.
Answer: Time.
==
Question: What is the name of the building?
Context: A white building is in front of the image. The building has
an arc roof. People are flying kites on the grassland.
Answer: White House.
==
Question: What city is this?
Context: A clock tower is standing next to Westminster Palace. Buses
are crossing the river along the bridge.
Answer: London
==
Question: What is the name of the tower by the River Thames?
```

```
Context: Ships are sailing by the River Thames. Big Ben Stands by the
shore of the River Thames.
Answer: Big Ben.
==
Question: Which country is famous for the dish?
Context: There are fried chips and potatoes on the plate. People are
having lunch outdoors by the roads. Answer: England.
==
Question: What is the tall building to the left of the boat?
Context: There is a tall building to the left of the boat. A boat is
parked by the land. There is a bridge crossing over the water. There
is a palace to the right of the tower.
Answer:

-- Returned --
Big Ben.
```

References

1. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (2017)
2. 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)
3. 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)
4. 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)
5. 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)
6. Gupta, N., Lin, K., Roth, D., Singh, S., Gardner, M.: Neural module networks for reasoning over text. arXiv preprint arXiv:1912.04971 (2019)
7. Honnibal, M., Montani, I.: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (2017), to appear
8. 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)
9. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM computing surveys (CSUR) **54**(6), 1–35 (2021)
10. 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)

11. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490 (2019)
12. Yang, X., Tang, K., Zhang, H., Cai, J.: Auto-encoding scene graphs for image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10685–10694 (2019)
13. 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)
14. Zhao, Z., Samel, K., Chen, B., et al.: Proto: Program-guided transformer for program-guided tasks. Advances in neural information processing systems **34**, 17021–17036 (2021)