Supplementary material for ”VQA-GNN: Reasoning with Multimodal Knowledge via Graph Neural Networks for Visual Question Answering”

A. Frequently asked questions (FAQ)

Q1: Does VQA-GNN performance depend on pretrained model accuracy? In general, the better the pre-trained model, the better performance we can achieve on downstream tasks. To eliminate the performance dependencies of the concept-graph and QA-concept node, our GQA evaluation builds the multimodal semantic graph with minimal components, QA context node, and (textual/visual) scene-graph. The results (Tab.5, Fig.5 of the main paper) show that VQA-GNN is effective in performing bidirectional fusion across structured and unstructured multimodal knowledge.

Q2: Does the whole model can be optimized end-to-end from the raw information? Currently, we do not train end-to-end as our model can be optimized well by using pretrained scene graph representations. However, our framework is general and we are extending our implementation to handle end-to-end training by performing multimodal semantic graph generation from raw images in a multi-task training setting.

Q3: Since GQA is built on VisualGenome, have you used it to build QA-concept node for GQA evaluation? No, the QA-concept node was not built in GQA evaluation.

Q4: How much of the performance gain is from these extra data and if it is a fair comparison with baselines? We conducted extensive ablation studies to understand the contributions of every piece of extra information (*i.e.*, scene-graph, concept-graph, QA-concept node and QA-context node [RoBERTa]) used for VCR evaluation: The results in Tab.2 of the main paper show that each extra information can be used for improving VQA-GNN performance. Specifically, the ”concept-graph + RoBERTa” can improve the performance of ”RoBERTa” by 15.2%, the ”scene-graph + concept-graph + RoBERTa” can improve 1.4%, and the ”scene-graph + concept-graph + RoBERTa + QA-concept” can further improve 2.0%.

We further provided a comparison result for the GQA dataset in Tab.4 of the main paper. By utilizing only the visual scene graph, VQA-GNN improved SGEITL by 5.6%. In addition, we evaluated the two technical innovations we made: multimodal GNN and bidirectional fusion (see Fig.4 and Fig.5 of the main paper). The results in Tab.3 show that VQA-GNN with the multimodal GNN on the unified multimodal semantic graph improved a single GNN by 2.1%. The results in Tab.5 show that VQA-GNN with the bidirectional fusion improved the unidirectional fusion by 4.0%.

Q5: Why was the baseline result reported instead of the better results in GraphVQA [1]? The better results reported in GraphVQA [1] use ground truth semantic function programs to provide powerful directives to the model, which is arguably a less realistic setting. To consider a more practical setting where the model needs to discover rational inference paths by itself (*e.g.*, from a unified multimodal graph; see §C), we did not use the semantic function programs, as in the baseline model GCN.

Q6: Why did you choose VCR and GQA for evaluating knowledge-based VQA? The VCR requires a wide range of commonsense knowledge, and the systems need not only pretrained unstructured knowledge (*e.g.*, QA-concept node and QA-context node), but also structured knowledge (*e.g.*, scene-graph and concept-graph). This motivated us to develop VQA-GNN that deeply and mutually fuses multimodal knowledge for visual question answering. Further, VQA-GNN can achieve strong compositional reasoning over textual and visual scene-graphs for GQA evaluation.

Q7: Do you have a plan to open source your project? Yes, we included the source code in the supplementary materials. We also plan to open source the code with preprocessed data after the review process.

B. Extra evaluation results

B.1. Comparison with baselines pretrained only on VCR dataset

We compared VQA-GNN and multimodal transformer models in Table 1 which were only trained on VCR dataset ($290K$ instances), as reported in SGEITL paper [2]. SGEITL is an add-on module that can boost multimodal transformer models (UNITER, VLBERT) by incorporating finetuned visual scene graph with multimodal transformer models. Compared with SGEITL, VQA-GNN is a GNN-based method built on the structured multimodal semantic graph. As shown in Table 1, VQA-GNN improves over SGEITL+VLBERT on the Q→AR metric by **4%** for the validation set, and further suggests the efficacy of VQA-GNN on the well-structured multimodal semantic graph.

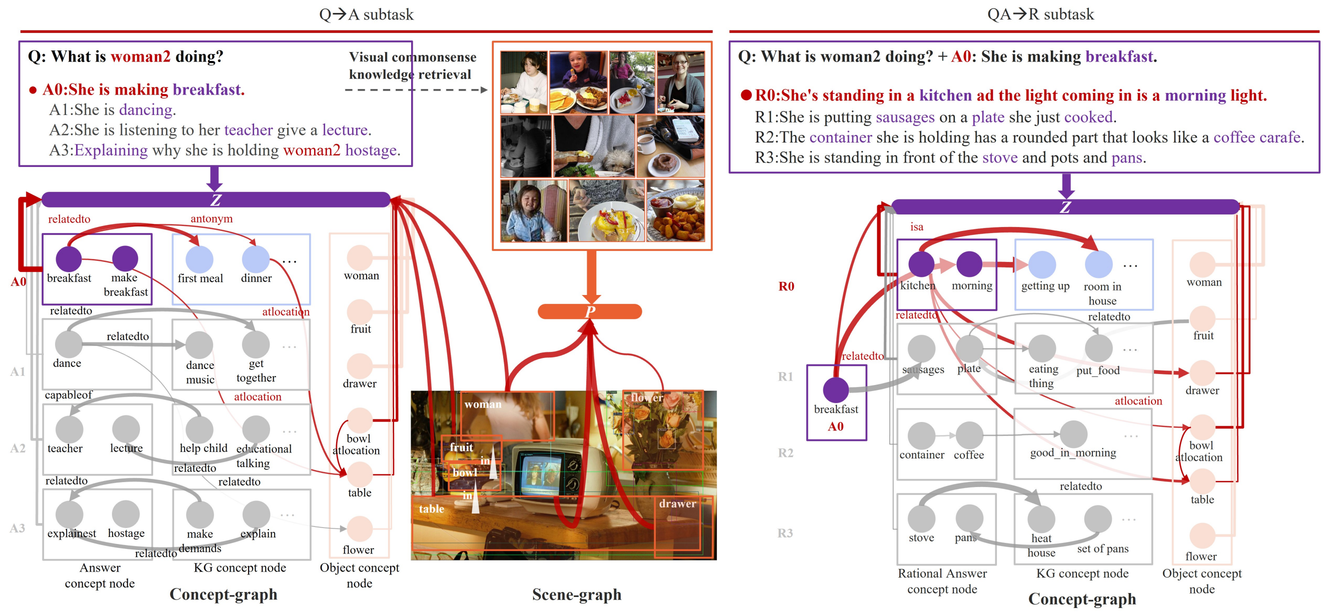

Figure 1. Interpreting *VQA-GNN*'s reasoning process across multimodal knowledge domains by indicating attention weight of the relationship between nodes. Arrows indicate the direction of the relationship, and darker and thicker edges indicate higher attention weights. The red color highlights the message passing routine for reasoning the correct answer and the gray color indicates the opposite.

Model	Parameters	Val Acc.(%)		
		Q→A	QA→R	Q→AR
VLBERT-L	383M	72.9	75.3	54.9
UNITER-L	378M	73.4	76.0	55.8
ERNIE-ViL-L	533M	74.1	76.9	56.9
SGEITL+UNITER	$>378M$	74.8	76.8	57.4
SGEITL+VLBERT	$>383M$	74.9	77.2	57.8
VQA-GNN (ours)	**372M**	**77.1**	**80.0**	**62.1**

Table 1. All models are trained only on the VCR dataset. Compared to ”SGEITL+VLBERT” model that inputs a visual scene graph to VLBERT network, VQA-GNN applied to a well-structured multimodal semantic graph improves accuracy on Q→AR metric by over **4%**.

Question type	Val Acc.(%) (Q→A)		Val Acc.(%) (QA→R)	
	VQA-GNN	RESERVE-L	VQA-GNN	RESERVE-L
Why	73.2	**78.6**	81.8	**84.8**
What	79.1	**85.7**	80.0	**85.2**
Where	77.9	**87.7**	76.7	**86.0**
Who	89.4	**91.3**	77.1	**85.0**
When	77.8	**94.4**	**100**	100
Which	**88.9**	88.9	81.5	**87.0**
Do	**81.6**	81.6	73.5	**82.5**
Will	**86.2**	83.8	**82.7**	82.3
Have	**92.9**	91.4	**87.1**	82.9
If	89.2	**92.3**	**96.9**	95.4
Can/Should	**93.3**	88.5	**87.5**	84.6

Table 2. Comparison results on the different question types. VQA-GNN performs better than RESERVE-L for ”Will”, ”Have” and ”Can/Should” question types.

B.2. Comparison results on different question types

We studied the performance of VQA-GNN in different question types and compared it with a strong baseline model RESERVE-L in Table 2. VQA-GNN outperforms RESERVE-L in some question types such as ”Will”, ”Have”, and ”Can/Should”, and we consider that some questions require the model to understand commonsense knowledge related to image context and have good reasoning ability. Hence, the model ”RESERVE-L+VQA-GNN” boosted the performance of RESERVE-L.

C. Interpretability

To interpret how *VQA-GNN* reason a correct answer based on a structured multimodal semantic graph, we show the reasoning process on Q→A and QA→R subtasks of VCR respectively in Figure 1 by using a validation sample that is given a correct answer on both Q→A and QA→R subtasks by *VQA-GNN*.

***Q→A* subtask.** We trace high attention weights from two directions: **d1**: QA-context node **Z** → **A**nswer nodes (purple) → **KG** concept nodes (blue) → **O**ject concept nodes (pink); **d2**: QA-concept node **P** → **SG** object nodes (orange) → **Z**. At the **d1**, **Z** pays more attention to **A** nodes ”breakfast” and ”make breakfast” in answer ”A0” choice than nodes in other choices, ”breakfast” attends to both **KG** node ”first meal” and **O** node ”table”, **O** node ”table” further attends to **O** node ”bowl”, and both strongly attend to **Z**. **A** node ”breakfast” bridges the reasoning between **Z** and **O** ”table” at the concept-level. Besides with **d1**, we also tract the attention weights from **d2**, **Z** strongly attends to **SG** nodes ”table”, ”drawer” and ”woman”, all nodes attend to **Z**, which reveals image-level semantic knowledge of **SG** nodes ”table”, ”drawer” and ”woman” are all essential for reasoning ”**A0:** she is making breakfast” correct. These two reasoning paths demonstrate that *VQA-GNN* is an inoperable method that can give a reasonable explanation to each choice with our well structured multimodal semantic graph, also suggest that our multimodal semantic graph is capable of unifying unstructured (*e.g.*, QA-context node and QA-concept node) and structured (*e.g.*, scene-graph and concept-graph) multimodal knowledge.

***QA*→*R* subtask.** We trace reasoning path for the rational answer **R0** on concept-graph. There are two reasonable directions: **Z** → ”breakfast” → ”morning” → ”getting up”, and **Z** → ”kitchen” → ”drawer”, ”bowl”, ”table”. Both of them show strong attentions between QA text and **R0**, compared to the attention direction for **R1** indicating that ”breakfast” also strongly attends to ”sausages” and ”plate” attends to ”fruit”, however, ”fruit” weakly attends to **Z**. As a result, *VQA-GNN* can select a rational answer, and suggest its interpretability on *QA*→*R* subtask. In addition, we noted that our method has close reasoning paths that attending to image context of ”bowl”, ”table” and ”drawer” on both *Q*→*A* and *QA*→*R* subtasks. Hence, we consider that our method has strong reasoning ability across multimodal knowledge domains.

Code is coming soon. https://github.com/yanan1989/VQA-GNN

References

[1] Weixin Liang, Yanhao Jiang, and Zixuan Liu. Graghvqa: Language-guided graph neural networks for graph-based visual question answering. *ArXiv*, abs/2104.10283, 2021. 1

[2] Zhecan Wang, Haoxuan You, Liunian Harold Li, Alireza Zareian, Suji Park, Yiqing Liang, Kai-Wei Chang, and Shih-Fu Chang. Sgeitl: Scene graph enhanced image-text learning for visual commonsense reasoning. In *AAAI*, 2022. 1