	$Q \rightarrow A$		$QA \rightarrow R$		$Q \rightarrow AR$	
Model	GloVe	BERT	GloVe	BERT	GloVe	BERT
BottomUpTopDown [2]	42.8	62.3	25.1	63.0	10.7	39.6
R2C [46]	46.4	63.8	38.3	67.2	18.3	43.1
HGL (ours)	**54.1**	**69.4**	**42.7**	**70.6**	**25.1**	**49.1**

Table 1: Results on the VCR validation set. While BERT helps the performance of these baselines, our model still performs the best.

To All: We thank all reviewers for the constructive and insightful comments and address the comments as follows. We appreciate the fact that all reviewers give positive evaluation results about our HGL. We promise to revise typos and grammatical mistakes and polish our paper in the revised version based on these valuable suggestions.

Q1: Clarification on results of Fig.4 and Fig.5 (R1). In Fig.4(a), "feeling" from question can be aligned with the most suitable semantic word "anger" from the answer for right answer prediction, which demonstrated the effectiveness of our QAHG. In Fig.4(e), the visual representation "person5" is aligned with the linguistic word "witness" correctly via VAHG, because "person5" is the "witness" in this scenario. Visual representation "person1" (tag associated with the red box is fed to model) is connected with the emotional word "angry" to achieve a heterogeneous relationship. In Fig.5, the person would fell wet because of rain in the scenario. The arrow is pointing at raindrop.

Q2: Comment on CVM applied to vision (R1). The CVM is suitable to apply to visual context, because there are more information can be obtained from the visual evidence. For instance, the first example of Figure 1 from supplementary material, the motivation of person1 must get information from visual evidence (e.g. cake1, table and background) instead of the question to predict the right answer and reason.

Q3: Discussions on the affection of concatenation in Eq.5 (R1). To combine the relationship between X_m and Y_o with Y_o instead of simply combining Y_o with X_m, we need the concatenation operation in Eq.5. We did experiments to analyze the affection and found the concatenation can improve ~0.4% accuracy (69 vs 69.4) on Q->A task over VCR validation set. The experiment showed that the concatenation operation can help the model to further understand the heterogeneous relationship for predicting correctly.

Q4: More experiments on common VQA datasets (R2). We trained our HGL on VQA2 dataset following the same experimental setup [2], and evaluated our HGL on VQA2 validation dataset compared with common VQA approach such as BottomUpTopDown [2]. The BottomUpTopDown [2] achieved 63.2 accuracy across all question types and the HGL achieved 65.3 accuracy, which showed our proposed model can also be transferred to VQA task and perform well.

Q5: Ablation experiments between BERT and common VQA backbone (e.g. GloVe) (R2). In Table 1, the BERT indeed helped the performance of HGL (54.1→69.4, 42.7→70.6, 25.1→49.1), our HGL still got SOTA performance compared with common VQA approach such as BottomUpTopDown [2], and R2C [46] on VCR validation set.

Q6: Clarification on CVM and the possibility to take "visual context" as graph nodes (R2). The CVM aims to capture ambiguous semantic context (e.g. rainy/snowy weather) that lack of specific labels for detection and can not benefit from the labeled object grounding boxes and categories such as "person" and "dog" during training. The long-range visual context means spacial visual evidence that is consisted of pixel-wise representations of different positions. It may not be suitable to take "visual context" as graph nodes. Because the "visual context" contains complex and uncertain semantics such as different weather and emotional expression, which is different to make sure each node semantics and the number of nodes to represent the whole visual context.

Q7: Clarification on graph learning, inputs and outpus of the model during training and inference (R3). The graphs are latent to learn. The inputs of our HGL are candidate answers, a question and an image. The output of the model is a prediction of right answers. The inputs and outputs of the model are the same during training and inference.

Q8: Explanation for a few questions on section 1 and section 3 (R1 R2 R3). In section 1, the concept of information isolated island in our paper refers to the independent of different semantic nodes can not achieve semantic inference in a homogeneous graph that connects similar semantic nodes by attribute (e.g. Figure 1(a)) or grammar (e.g. Figure 1(b)). The y_i^{rl} should be y_i^l in Eq.12, we will revise the typos in final version. In L131, the correlations between the vectors are computed via a dot product, we promise to give a more precise definition of this operation in the final version. In section 3.1, maybe "Graph Construction" would be better than the title "Task Definition", we also realize the inappropriate title and will consider carefully the name of the title in our final version. We will briefly describe the three tasks (Q->A, QA->R and Q->AR) and some specific terminologies in our final version to make the paper easier to understand.

Q9: More qualitative analysis and the application of this work (R3). Due to the space limit, we will show more qualitative results with analysis in our final version. Our HGL can seamlessly integrate the intra-graph and inter-graph reasoning in order to bridge vision and language domain and can interactively refine reasoning paths for semantic agreement. Such capability would show more potentials in facilitating high-level applications that require cross-domain semantic alignment among visual concepts and linguistic words. (e.g. visual grounding and VQA).