Table 1: Evaluation of the state-of-the-art model SGAE on the image captioning dataset.

Model	B@4	M	R	C	S
SGAE	38.8	28.5	58.7	129.3	22.3
w/ MIA	**39.6**	**29.0**	**58.9**	**130.1**	**22.8**

Table 2: The accuracy on the VQA v2.0 test set.

Model	Number	Other	Overall	Test-std
Up-Down	44.2	56.2	65.6	65.9
w/ MIA	**51.2**	**59.7**	**68.8**	**69.1**
BAN	50.9	60.3	69.6	69.8
w/ MIA	**53.1**	**60.5**	**70.2**	**70.3**

We thank all the reviewers for the helpful comments. As we replied, we will revise the paper to address the concerns and further proofread the paper.

Q1: How the paper's contribution relates to the current SOTA?

A1: For image captioning, we simply had not tried to combine MIA with SGAE [2], because our main goal is to show that MIA could improve the performance of a wide range of established models. SGAE is a rather complicated scene-graph based method specific to image captioning. But it is indeed interesting to see whether MIA can still help SGAE since scene-graphs are also finer representations of an image. As presented in Table 1, it is clear that MIA also boosts the performance of SGAE (a higher new SOTA), indicating that MIA learns very effective representations even for scene-graphs.

For VQA, as shown in Table 2, the two state-of-the-art systems for VQA also achieve an overall improvement on the test set, which is in accordance with the results on the validation set in the paper.

The results with current SOTA + MIA will be stated more clearly in the paper.

Q2: How to use MIA on the baseline systems (i.e., how is MIA applied to image captioning where the inputs are only images) and what are the experimental settings?

A2: We can extract visual and textual features (see Figure 1 in the paper) for an image, regardless of the tasks (image captioning or VQA). This indicates that MIA does not require original text information about the image (e.g., the questions in VQA), which is not provided in the task of image captioning. For using the semantic-grounded features learned by MIA, we can simply replace the original source features with MIA-refined features. Specifically, in Figure 3 of the paper, traditional LSTM-A3 takes as input textual features at the first decoding step and visual features at the second step. Since MIA will not change the number or the size of the feature vectors (each of them can be seen as a weighted average of the original features), we can replace the original features directly.

For the settings, we have listed them in the supplementary materials. In particular, we preserve the original settings for all the baseline models, since our focus is to provide better image representations.

Q3: How is the textual concept extracted and processed?

A3: We predict textual concepts using the textual concept extractor proposed by Fang et al. (2015) [1], which is based on fully convolutional network (FCN) and is purely trained on the training data of COCO image captioning dataset for concept prediction, i.e., we do not adopt the ground-truth textual concepts (if the ground-truth textual concepts are utilized, the SGAE w/ MIA model can achieve a CIDEr score of 168.7, which indicates that better textual concepts will give rise to better performance of downstream tasks). As shown in line 83 of the paper, we simply apply word embedding (randomly initialized), which is shared with the caption/question words, to process textual concepts. It is worth noticing that in our baselines, no further processing except MIA is done. Besides, we do not use extra dataset to train MIA. The MIA is trained jointly with the baseline models from scratch, therefore no external knowledge is introduced to boost the performance. Especially, as presented in Table 3 of the paper, when applied to LSTM-A3 and LSTM-A4, which use the same (visual and textual) features as we adopted, MIA can still boost the performance, further demonstrating the effectiveness of alignment between visual regions and textual concepts.

References

[1] H. Fang, S. Gupta, F. N. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. From captions to visual concepts and back. In *CVPR*, 2015.

[2] X. Yang, K. Tang, H. Zhang, and J. Cai. Auto-encoding scene graphs for image captioning. In *CVPR*, 2019.