TABLE I
AUTOMATIC EVALUATION RESULTS

Dataset	BLEU 1	BLEU 2	BLEU 3	BLEU 4	ROUGE-L
MMQA	34.08	23.90	17.32	13.32	22.13
MMQA (Text-only)	35.67	24.26	18.34	14.62	23.29
SQuAD	46.97	33.72	24.28	18.89	31.24
MMQA (Model fine-tuned using SQuAD)	37.37	27.54	21.02	16.05	28.27

may be differences in the way that BLEU scores are calculated by different libraries [18].

TABLE II
HUMAN EVALUATION RESULT

Modality	Fluency	Relevance
Image and Text	55%	53%
Text	76%	78%

To evaluate the results from the model outputs, a human evaluation was conducted, especially the intrinsic human evaluation method, which evaluates the questions in terms of their fluency and relevance [18]. A fluent question is a question stated in a clear and confident manner without any hesitation or unnatural pauses [19]. While a relevant question is a question related to the scope of context being discussed [20]. We conduct the intrinsic human evaluation in form of a questionnaire given to humans. To measure the fluency and relevance of the generated questions, a 5-point Likert scale was used to measure 50 random samples from all generated questions in the test set, inspired by a human evaluation conducted in [21].

Results from Table II show that this study succeeds in conducting the intrinsic human evaluation. From the result, we can conclude that questions generated from multimodal reasoning (text and image) affect their fluency and relevance, while the generated questions from text-only modality have more quality in terms of fluency and relevance. There are an increase of 38% in fluency and 47% in relevance when the generated questions are from a text-only modality. Although the results of the text-only modality are higher than those using the image and text modalities, the proposed system is still able to generate questions with quite a good result (above 50% in terms of fluency and relevance).

V. CONCLUSION

We can conclude that this study succeeds to build a model to tackle the Multimodal Question Generation task using the combination of MAG and BERT-based model called MAG-BERT. All approaches from this study show that the model needs to be fine-tuned first using more simpler dataset and more training data like SQuAD. It's proven when the model is fine-tuned first using SQuAD to get used to generate questions, the model has more quality results when fine-tuned again using the MMQA dataset. The last approach resulted in a 16.05 BLEU 4 and 28.27 ROUGE-L score, which is the highest among other approaches used when using the MMQA dataset.

To improve the quality of generated questions from the proposed model, we need to enhance the model performance when dealing with more than one modality like an image. Future work on image context could contribute to creating a better image representation so that the model understands the connection between the image and text context. In the upcoming study, an approach of sequence-to-sequence model such as BART could be considered to enhance the quality of the generated questions.

REFERENCES

[1] Y. Zhao, X. Ni, Y. Ding, and Q. Ke, "Paragraph-level neural question generation with maxout pointer and gated self-attention networks," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 3901–3910. [Online]. Available: https://aclanthology.org/D18-1424

[2] Y.-H. Chan and Y.-C. Fan, "BERT for question generation," in *Proceedings of the 12th International Conference on Natural Language Generation*. Tokyo, Japan: Association for Computational Linguistics, Oct.–Nov. 2019, pp. 173–177. [Online]. Available: https://aclanthology.org/W19-8624

[3] D. Xiao, H. Zhang, Y. Li, Y. Sun, H. Tian, H. Wu, and H. Wang, "ERNIE-GEN: an enhanced multi-flow pre-training and fine-tuning framework for natural language generation," *CoRR*, vol. abs/2001.11314, 2020. [Online]. Available: https://arxiv.org/abs/2001.11314

[4] N. Duan, D. Tang, P. Chen, and M. Zhou, "Question generation for question answering," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 866–874. [Online]. Available: https://aclanthology.org/D17-1090

[5] B. N. Patro, S. Kumar, V. K. Kurmi, and V. Namboodiri, "Multimodal differential network for visual question generation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 4002–4012. [Online]. Available: https://aclanthology.org/D18-1434

[6] W. Rahman, M. K. Hasan, S. Lee, A. Bagher Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 2359–2369. [Online]. Available: https://aclanthology.org/2020.acl-main.214

[7] Y.-H. Chan and Y.-C. Fan, "A recurrent BERT-based model for question generation," in *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 154–162. [Online]. Available: https://aclanthology.org/D19-5821

[8] A. Talmor, O. Yoran, A. Catav, D. Lahav, Y. Wang, A. Asai, G. Ilharco, H. Hajishirzi, and J. Berant, "Multimodal{qa}: complex question answering over text, tables and images," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=ee6W5UgQLa

[9] Q. Zhou, N. Yang, F. Wei, C. Tan, H. Bao, and M. Zhou, "Neural question generation from text: A preliminary study," *CoRR*, vol. abs/1704.01792, 2017. [Online]. Available: http://arxiv.org/abs/1704.01792

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008. [Online]. Available: http://arxiv.org/abs/1706.03762

[11] X. Lin, G. Bertasius, J. Wang, S. Chang, D. Parikh, and L. Torresani, "VX2TEXT: end-to-end learning of video-based text generation from multimodal inputs," *CoRR*, vol. abs/2101.12059, 2021. [Online]. Available: https://arxiv.org/abs/2101.12059

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[13] F. Alfaro, M. R. Costa-jussà, and J. A. R. Fonollosa, "BERT masked language modeling for co-reference resolution," in *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 76–81. [Online]. Available: https://aclanthology.org/W19-3811

[14] R. Zhang, J. Guo, L. Chen, Y. Fan, and X. Cheng, "A review on question generation from natural language text," *ACM Trans. Inf. Syst.*, vol. 40, no. 1, sep 2021. [Online]. Available: https://doi.org/10.1145/3468889

[15] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040

[16] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[17] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2383–2392. [Online]. Available: https://aclanthology.org/D16-1264

[18] C. van der Lee, A. Gatt, E. van Miltenburg, S. Wubben, and E. Krahmer, "Best practices for the human evaluation of automatically generated text," in *Proceedings of the 12th International Conference on Natural Language Generation*. Tokyo, Japan: Association for Computational Linguistics, Oct.–Nov. 2019, pp. 355–368. [Online]. Available: https://aclanthology.org/W19-8643

[19] I. J. O. T. Education, "Fluency as successful communication," 12 2018.

[20] J. Mason and T. Hoel, "The relevant question and the question of relevance," 01 2011.

[21] L. Pan, W. Lei, T. Chua, and M. Kan, "Recent advances in neural question generation," *CoRR*, vol. abs/1905.08949, 2019. [Online]. Available: http://arxiv.org/abs/1905.08949