

Supplementary Materials: AutoGraph: Enabling Visual Context via Graph Alignment in Open-Domain Multi-Modal Dialogue Generation

Deji Zhao*	Donghong Han†	Ye Yuan	Bo Ning
Northeastern University	Northeastern University	Beijing Institute of Technology	Dalian Maritime University
School of Computer Science and Engineering	School of Computer Science and Engineering	School of Computer Science and Technology	School of Information Science and Technology
Shenyang, China	Shenyang, China	Beijing, China	Dalian, China
zhaodeji@stumail.neu.edu.cn	handonghong@cse.neu.edu.cn	yuan-ye@bit.edu.cn	ningbo@dlmu.edu.cn
Mengxiang Li	Zhongjiang He	Shuangyong Song	
China Telecom Corp Ltd	China Telecom Corp Ltd	China Telecom Corp Ltd	
Beijing, China	Institute of Artificial Intelligence (TeleAI)	Institute of Artificial Intelligence (TeleAI)	
limengx@126.com	Beijing, China	Beijing, China	
	hezj@chinatelecom.cn	songsy@chinatelecom.cn	

LIMITATIONS

The AutoGraph method is a general approach, where its internal components can be easily replaced, such as the extraction of scene graphs and text graphs. At present, there are numerous methods for extracting scene graphs and dependency graphs, and we don't test all of them. With the development of technology, the quality of automatically constructed visual context graph is expected to improve. We also believe that visual context graphs can potentially include more information, such as emotions of characters, relationships between entities, and information inferred from these relationships, which are not considered in this paper.

Furthermore, graph semantic alignment is a little time-consuming. In future work, we will explore to enhance the speed of automated graph construction. And there are various large language models available, but due to computational constraints, we chose to use widely recognized models such as Llama2-7B [4] and LLaVA-7B [1] as our base models.

ETHICAL STATEMENT

MELD dataset [3] and OpenViDial dataset [2] providers filter all personal information and obscene language. We believe that the dataset used in our experiments is harmless to users. If applied to the real world, it is necessary to consider the security of the model and avoid responding to harmful and biased responses.

The participants in our human evaluation are volunteered transparently informed of our research intent, with reasonable wages paid.

REFERENCES

- [1] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems* 36 (2024).

- [2] Yuxian Meng, Shuhe Wang, Qinghong Han, Xiaofei Sun, Fei Wu, Rui Yan, and Jiwei Li. 2020. Openvidial: A large-scale, open-domain dialogue dataset with visual contexts. *arXiv preprint arXiv:2012.15015* (2020).
- [3] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 527–536.
- [4] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jianxiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *CoRR* abs/2307.09288 (2023). <https://doi.org/10.48550/ARXIV.2307.09288> arXiv:2307.09288

*Work done during internship at TeleAI.

†Corresponding author.