LGM3A 2024

The 2nd Workshop on Large Generative Models Meet Multimodal Applications

Xu, Shihao; Luo, Yiyang; Dauwels, Justin; Khong, Andy; Wang, Zheng; Chen, Qianqian; Cai, Chen; Shi, Wei; Chua, Tat Seng

DOI
10.1145/3688866.3696056

Publication date
2024

Document Version
Final published version

Published in
LGM3A '24

LGM^{3}A 2024: the 2nd Workshop on Large Generative Models Meet Multimodal Applications

Shihao Xu
Huawei Singapore Research Center
Singapore, Singapore
shihao.xu@huawei.com

Yiyang Luo
Huawei Singapore Research Center
Singapore, Singapore
luoyiyang2@huawei.com

Justin Dauwels
Delft University of Technology
Delft, the Netherlands
j.h.g.dauwels@tudelft.nl

Andy Khong
Nanyang Technological University
Singapore, Singapore
andykhong@ntu.edu.sg

Zheng Wang
Huawei Singapore Research Center
Singapore, Singapore
wangzheng155@huawei.com

Qianqian Chen
Huawei Singapore Research Center
Singapore, Singapore
chenqianqian20@huawei.com

Chen Cai
Huawei Singapore Research Center
Singapore, Singapore
cai.chen2@huawei.com

Wei Shi
Huawei Singapore Research Center
Singapore, Singapore
w.shi@huawei.com

Tat-Seng Chua
National University of Singapore
Singapore, Singapore
chuats@comp.nus.edu.sg

Abstract

This workshop aims to explore the potential of large generative models to revolutionize how we interact with multimodal information. A Large Language Model (LLM) represents a sophisticated form of artificial intelligence engineered to comprehend and produce natural language text, exemplified by technologies such as GPT, LLaMA, Flan-T5, ChatGLM, Qwen, etc. These models undergo training on extensive text datasets, exhibiting commendable attributes including robust language generation, zero-shot transfer capabilities, and In-Context Learning (ICL). With the surge in multimodal content—encompassing images, videos, audio, and 3D models—over the recent period, Large MultiModal Models (LMMs) have seen significant enhancements. These improvements enable the augmentation of conventional LLMs to accommodate multimodal inputs or outputs, as seen in BLIP, Flamingo, KOSMOS, LLaVA, Gemini, GPT-4, etc. Concurrently, certain research initiatives have developed specific modalities, with Kosmos2 and MiniGPT-5 focusing on image generation, and SpeechGPT on speech production. There are also endeavors to integrate LLMs with external tools to achieve a near "any-to-any" multimodal comprehension and generation capacity, illustrated by projects like Visual-ChatGPT, ViperGPT, MMREACT, HuggingGPT, and AudioGPT. Collectively, these models, spanning not only text and image generation but also other modalities, are referred to as large generative models. This workshop will allow researchers, practitioners, and industry professionals to explore the latest trends and best practices in the multimodal applications of large generative models.

LGM3A '24, October 28-November 1 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-1193-0/24/10
https://doi.org/10.1145/3688866.3696056

CCS Concepts

• **Information systems** → **Multimedia information systems**.

Keywords

large language models, generative models, multimodal applications

ACM Reference Format:
Shihao Xu, Yiyang Luo, Justin Dauwels, Andy Khong, Zheng Wang, Qianqian Chen, Chen Cai, Wei Shi, and Tat-Seng Chua. 2024. LGM^{3}A 2024: the 2nd Workshop on Large Generative Models Meet Multimodal Applications. In *Proceedings of the 2nd Workshop on Large Generative Models Meet Multimodal Applications (LGM3A '24), October 28-November 1 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3688866.3696056

1 Introduction

The cross-modal generation has achieved significant progress in recent years. With a combination of multiple modalities (e.g., image, text, audio, etc.), multimodal methods achieve state-of-the-art performance not only on the cross-modality tasks, but also on the vision and NLP tasks. However, how to combine the current large pretraining models with the multimodal data to improve the performance of the user-engaged tasks is still to be explored.

The workshop's focus on multimodal generation and analysis, and the integration of different forms of multimedia information, is a topic of interest for a wide range of communities, including computer vision, multimedia, artificial intelligence, human-computer interaction, and others. Multimodal applications on large generative models have many potentials uses in various scenarios including visual question answering, text-to-image synthesis, speech-to-text synthesis and data augmentation which could interest many IT companies such as Google, Microsoft, TikTok, Baidu, Alibaba, Tencent, etc. In summary, the 2nd Workshop on Large Generative Model Meets Multimodal Applications workshop is relevant to the ACM Multimedia community, it addresses a critical area of research

within natural language understanding and computer vision, making it an important and timely event for researchers, practitioners, and students in the field.

2 Scope and Topics of The Workshops

The workshop will cover a wide range of topics including but not limited to:

- Multimodal content creation
- Multimodal data analysis and understanding
- Multimodal question answering
- Multimodal information retrieval
- Multimodal recommendation
- Multimodal summarization and text generation
- Multimodal conversational agents
- Multimodal machine translation
- Multimodal fusion and integration of information
- Multimodal applications/pipelines
- Multimodal systems management and indexing

The workshop will also focus on exploring the challenges and opportunities of integrating large language models with other AI technologies such as computer vision and speech recognition. It provides a platform for participants to present their research, share their experiences and discuss potential collaborations.

3 Relationship to previous workshops

The first LGM3A workshop was held successfully at ACM MM 2023 [2], with about 20 submissions and 8 high-quality papers accepted. We also invited three keynote speakers: Prof. Ziwei Liu, Prof. Boyang Li, and Prof. Zheng Shou to give talks, attracting many researchers to the workshop. The workshop on multimodal applications of large language models offers a unique perspective on the combination of language, vision, and audio and their applications. It provides a platform for presenting cutting-edge research and discussing future directions in this emerging field.

4 participants and invited speakers

Ziwei Liu
Affiliation: Nanyang Technological University
Biography: Ziwei Liu is currently a Nanyang Assistant Professor at Nanyang Technological University, Singapore. His research revolves around computer vision, machine learning and computer graphics. He has published extensively on top-tier conferences and journals in relevant fields, including CVPR, ICCV, ECCV, NeurIPS, ICLR, ICML, TPAMI, TOG and Nature - Machine Intelligence. He is the recipient of Microsoft Young Fellowship, Hong Kong PhD Fellowship, ICCV Young Researcher Award, HKSTP Best Paper Award and WAIC Yunfan Award. He serves as an Area Chair of CVPR, ICCV, NeurIPS and ICLR, as well as an Associate Editor of IJCV.

Zheng Shou
Affiliation: National University of Singapore
Biography: Zheng Shou is a tenure-track Assistant Professor at National University of Singapore. He was a Research Scientist at Facebook AI in Bay Area. He obtained his Ph.D. degree at Columbia University in the City of New York, working with Prof. Shih-Fu Chang. He was awarded Wei Family Private Foundation Fellowship. He received the best paper finalist at CVPR'22, the best student paper nomination at CVPR'17. His team won the 1st place in the international challenges including ActivityNet 2017, Ego4D 2022, EPIC-Kitchens 2022. He is a Fellow of National Research Foundation (NRF) Singapore. He is on the Forbes 30 Under 30 Asia list.

5 Workshop Organizers

Shihao Xu is a Research Scientist at Huawei Singapore Research Center, a multimodal search and recommendation lab. His current research interests and works fill in multimodal applications including sports video representations, user intention generation, multimodal geometry problem solving, and multimodal prompting. He received his PhD degree in Nanyang Technological University in 2022, advised by Prof. Justin Dauwels and Prof. Andy Khong. He received his Master's degree from Nanyang Technological University and Bachelor's degree from Harbin Institute of Technology. During his Ph.D., he was working on the audio-visual understanding of human behaviors.

Yiyang Luo is currently a Multimodal Search Algorithm Engineer at Huawei Singapore Research Centre, Multimodal Search and Recommendation Lab. He received his Master's degree from Nanyang Technological University and his Bachelor's degree from the Chinese University of Hong Kong. His research interests include multimodal deep learning and prompt engineering.

Justin Dauwels starts in January 2021 as an Associate Professor at TU Delft. Before this, he was an Associate Professor with the School of Electrical Electronic Engineering at Nanyang Technological University (NTU), Singapore. He obtained a PhD degree in electrical engineering at the Swiss Polytechnical Institute of Technology (ETH) in Zurich in December 2005. Next, from 2006-to 2007 he was a postdoc at the RIKEN Brain Science Institute, Japan (Prof. Shunichi Amari and Prof. Andrzej Cichocki), and a research scientist during 2008-2010 in the Stochastic Systems Group (SSG) at the Massachusetts Institute of Technology (MIT), led by Prof. Alan Willsky. His research interests are in data analytics with applications to intelligent transportation systems, autonomous systems, and analysis of human behavior and physiology. He obtained his PhD degree in electrical engineering at the Swiss Polytechnical Institute of Technology (ETH) in Zurich in December 2005. Moreover, he was a postdoctoral fellow at the RIKEN Brain Science Institute (2006-2007) and a research scientist at the Massachusetts Institute of Technology (2008-2010). He has been elected as an IEEE SPS 2024 Distinguished Lecturer. He has been a JSPS postdoctoral fellow (2007), a BAEF fellow (2008), a Henri-Benedictus Fellow of the King Baudouin Foundation (2008), and a JSPS invited fellow (2010, 2011). He served as Chairman of the IEEE CIS Chapter in Singapore from 2018 to 2020. He served as Associate Editor of the IEEE Transactions on Signal Processing (2018 - 2023), Associate Editor of the Elsevier journal Signal Processing (since 2021), member of the Editorial Advisory Board of the International Journal of Neural Systems, and organizer of IEEE conferences and special sessions. He was also Elected Member of the IEEE Signal Processing Theory and Methods Technical Committee and IEEE Biomedical Signal Processing Technical Committee (2018-2023).

Andy Khong is currently an Associate Professor in the School of Electrical and Electronic Engineering, at Nanyang Technological University, Singapore. Before that, he obtained his Ph.D. ('02-'05) from the Department of Electrical and Electronic Engineering, Imperial College London, after which he also served as a research associate ('05-'08) in the same department. He obtained his B.Eng. ('98-'02) at Nanyang Technological University in Singapore. His postdoctoral research involved the development of signal processing algorithms for vehicle destination inference as well as the design and implementation of acoustic array and seismic fusion algorithms for perimeter security systems. His Ph.D. research was mainly on partial-update and selective-tap adaptive algorithms with applications to mono- and multi-channel acoustic echo cancellation for hands-free telephony. He has also published works on speech enhancement, multi-channel microphone array, and blind deconvolution algorithms. His other research interests include education data mining, and machine learning applied to education data. Andy currently serves as an Associate Editor for the IEEE Trans. Audio, Speech and Language Processing and the Journal of Multidimensional Systems and Signal Processing (Springer). He was a visiting professor at UIUC in 2012 under the Tan Chin Tuan Fellowship. He is the author and co-author of two papers awarded the "Best Student Paper Awards" and is a recipient of the Junior Chambers International "Ten Outstanding Young Persons Honor Award 2011" and the Institute of Singapore "Prestigious Engineering Achievement Award 2012." He was awarded the Nanyang Education Award and the Educator of the Year Award in 2022.

Zheng Wang is currently a Principal Researcher and Huawei TopMinds at Huawei Singapore Research Center. His current research interest focuses on multimodal content generation and search. Before that, he received his PhD degree at the School of Computer Science and Engineering, Nanyang Technological University in 2022, advised by Prof. Cheng Long and Prof. Gao Cong. He received his Master's degree from the Department of Computer Science, the University of Hong Kong in 2018, and his Bachelor's degree from the School of Computer Science and Technology (Elite Class), Shandong University in 2016. Up to now, he has published over 20 papers in top conferences and journals, including SIGMOD, VLDB, ICDE, KDD, WWW, ACL, AAAI, and TKDE. Among them, his work MMQS [1] has been transferred to products, which indicates its significant impacts on both industry and academia. His research has been recognized by many prestigious awards, including Nominated Schmidt Science Fellows in 2023, World Artificial Intelligence Conference (WAIC) Yunfan Award Finalist in 2022, Google PhD Fellowship (sole winner from Asia in Database Management) in 2021, and AISG PhD Fellowship in 2021 (one of top three NTU awardees). He is also nominated for the NTU Best Thesis Award 2023 (under evaluation). He serves as a PC member (reviewer) for some top-tier conferences and journals, including KDD, NeurIPS, AAAI, CIKM, OSDI (Reproducibility), ATC (Reproducibility), DASFAA and TKDE.

Qianqian Chen is currently a Multimodal Search Algorithm Engineer at Huawei Singapore Research Centre, Multimodal Search and Recommendation Lab. She received her MSc Degree from Nanyang Technological University and his BSc Degree from Central South University. Her research interests include multimodal deep learning and prompt engineering.

Chen Cai is currently a Multimodal Search Algorithm Engineer at Huawei Singapore Research Centre, Multimodal Search and Recommendation Lab. He received his PhD Degree from Nanyang Technological University. His research interests include multimodal deep learning and prompt engineering.

Wei Shi is currently head of multimodal search team at Huawei Singapore Research Center. He received his PhD degree at Department of Computer Science and Technology, Tsinghua University in 2015. His research interests are broadly in multimodal search, vision-language alignment, and big data systems.

Tat-Seng Chua is the KITHCT Chair Professor at the School of Computing, National University of Singapore (NUS). He is also the Distinguished Visiting Professor of Tsinghua University, the Visiting Pao Yue-Kong Chair Professor of Zhejiang University, and the Distinguished Visiting Professor of Sichuan University. Dr. Chua was the Founding Dean of the School of Computing from 1998-2000. His main research interests include unstructured data analytics, video analytics, conversational search and recommendation, and robust and trustable AI. He is the Co-Director of NExT, a joint research Center between NUS and Tsinghua University, and Sea-NExT, a joint Lab between Sea Group and NExT. Dr. Chua is the recipient of the 2015 ACM SIGMM Achievements Award, and the winner of the 2022 NUS Research Recognition Award. He is the Chair of steering committee of Multimedia Modeling (MMM) conference series, and ACM International Conference on Multimedia Retrieval (ICMR) (2015-2018). He is the General Co-Chair of ACM Multimedia 2005, ACM SIGIR 2008, ACM Web Science 2015, ACM MM-Asia 2020, and the upcoming ACM conferences on WSDM 2023 and TheWebConf 2024. He serves in the editorial boards of three international journals. Dr. Chua is the co-Founder of two technology startup companies in Singapore. He holds a PhD from the University of Leeds, UK.

6 Program Committee

We appreciate the reviewers' efforts and would like to thank the members of the PC for their valuable support: **Jieer Ouyang** (Huawei Singapore Research Center), **Bingzheng Gan** (Huawei Singapore Research Center), **Tianyi Zhang** (Huawei Singapore Research Center), **Teo Shu Xian** (Huawei Singapore Research Center)

7 Workshop Statistics

We would like to thank the ACM MM'24 conference organizers for agreeing to host our workshop and for their support, and all reviewers for their time and helpful contributions. The workshop in its first edition attracted 10 submissions, where 5 were accepted for publication. In addition, we invite three keynote speakers to present their original research in this field.

References

[1] Zheng Wang, Bingzheng Gan, and Wei Shi. 2024. Multimodal query suggestion with multi-agent reinforcement learning from human feedback. In *Proceedings of the ACM on Web Conference 2024*. 1374–1385.

[2] Zheng Wang, Cheng Long, Shihao Xu, Bingzheng Gan, Wei Shi, Zhao Cao, and Tat-Seng Chua. 2023. LGM3A'23: 1st Workshop on Large Generative Models Meet Multimodal Applications. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9744–9745.