LGM^{3}A 2024 Chairs’ Welcome

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

Publication date
2024

Document Version
Final published version

Published in
LGM³A ’24 Proceedings of the 2nd Workshop on Large Generative Models Meet Multimodal Applications

Association for
Computing Machinery

Advancing Computing as a Science & Profession

October 28–November 1, 2024
Melbourne, VIC, Australia

LGM^{3}A '24

Proceedings of the 2nd Workshop on

Large Generative Models Meet Multimodal Applications

Sponsored by:

ACM SIGMM

Organizers:

Shihao Xu (Huawei Singapore Research Center)
Yiyang Luo (Huawei Singapore Research Center)
Justin Dauwels (Delft University of Technology)
Andy Khong (Nanyang Technological University)
Zheng Wang (Huawei Singapore Research Center)
Qianqian Chen (Huawei Singapore Research Center)
Chen Cai (Huawei Singapore Research Center)
Wei Shi (Huawei Singapore Research Center)
Tat-Seng Chua (National University of Singapore)

Association for
Computing Machinery

Advancing Computing as a Science & Profession

The Association for Computing Machinery
1601 Broadway, 10th Floor
New York, NY 10019-7434

ISBN: 979-8-4007-1193-0

Additional copies may be ordered prepaid from:

ACM Order Department
PO Box 30777
New York, NY 10087-0777, USA

Phone: 1-800-342-6626 (USA and Canada)
+1-212-626-0500 (Global)
Fax: +1-212-944-1318
E-mail: acmhelp@acm.org
Hours of Operation: 8:30 am – 4:30 pm ET

Cover photo obtained from bigstockphoto.com

LGM^{3}A 2024 Chairs' Welcome

On behalf of the organizing committee, it is our distinct pleasure to extend a warm welcome to the LGM^{3}A Workshop. As Chairs of this conference, we are delighted to bring together a community of scholars, researchers, and professionals from diverse backgrounds, all driven by a shared passion for advancing the frontiers of knowledge in our field.

This workshop aims to explore the potential of large generative models to revolutionize the way 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, and 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 delved into generating 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 provide an opportunity for researchers, practitioners, and industry professionals to explore the latest trends and best practices in the field of multimodal applications of large generative models. We also remark that the submissions are not limited to the use of such models. 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. Additionally, the workshop will provide a platform for participants to present their research, share their experiences, and discuss potential collaborations.

We extend our sincere thanks to the members of the organizing committee, whose dedication and tireless efforts have brought this event to fruition. With their help, we have assembled 5 strong papers and 2 invited talks that will be presented at the conference. This is around 50% acceptance rate for regular papers. We also express our gratitude to our sponsors and partners for their invaluable support in making this conference possible.

Welcome to the LGM^{3}A Workshop, and let us collectively chart the path to new horizons in multimodal applications and large generative models.

Shihao Xu
Huawei Singapore Research Center

Yiyang Luo
Huawei Singapore Research Center

Justin Dauwels
Delft University of Technology

Andy Khong
Nanyang Technological University

Zheng Wang
Huawei Singapore Research Center

Qianqian Chen
Huawei Singapore Research Center

Chen Cai
Huawei Singapore Research Center

Wei Shi
Huawei Singapore Research Center

Tat-Seng Chua
National University of Singapore

Table of Contents

LGM^{3}A 2024 Workshop Organization

Organizers:
Shihao Xu (Huawei Singapore Research Center)
Yiyang Luo (Huawei Singapore Research Center)
Justin Dauwels (Delft University of Technology)
Andy Khong (Nanyang Technological University)
Zheng Wang (Huawei Singapore Research Center)
Qianqian Chen (Huawei Singapore Research Center)
Chen Cai (Huawei Singapore Research Center)
Wei Shi (Huawei Singapore Research Center)
Tat-Seng Chua (National University of Singapore)

Program Committee:
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)

Sponsor: SIGMM

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.

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

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

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.

Multi-Modal Generative AI with Foundation Models

Ziwei Liu
Nanyang Technological University
Singapore
ziwei.liu@ntu.edu.sg

ABSTRACT

Generating photorealistic and controllable visual contents has been a long-pursuing goal of artificial intelligence (AI), with extensive real-world applications. It is also at the core of embodied intelligence. In this talk, I will discuss our work in AI-driven visual context generation of humans [1, 2], objects [3] and scenes [4], with an emphasis on combining the power of neural rendering with large multimodal foundation models [5]. Our generative AI framework has shown its effectiveness and generalizability on a wide range of tasks.

CCS Concepts/ACM Classifiers

• Computing Methodologies → Artificial Intelligence → Computer Vision

Author Keywords

Computer vision; deep learning; generative AI; multimodal learning; foundation models

BIOGRAPHY

Prof. 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, SIGGRAPH, TPAMI, TOG and Nature - Machine Intelligence, with around 30,000 citations. He is the recipient of Microsoft Young Fellowship, Hong Kong PhD Fellowship, ICCV Young Researcher Award, HKSTP Best Paper Award, CVPR Best Paper Award Candidate, WAIC Yunfan Award and ICBS Frontiers of Science Award. He has won the championship in major computer vision competitions, including DAVIS Video Segmentation Challenge 2017, MSCOCO Instance Segmentation Challenge 2018, FAIR Self-Supervision Challenge 2019, Video Virtual Try-on Challenge 2020 and Computer Vision in the Wild Challenge 2022. He is also the lead contributor of several renowned computer vision benchmarks and softwares, including CelebA, DeepFashion, MMHuman3D and MMFashion. He serves as an Area Chair of CVPR, ICCV, NeurIPS and ICLR, as well as an Associate Editor of IJCV.

REFERENCES

[1] Text2Human: Text-Driven Controllable Human Image Generation. *TOG 2022.*

[2] AvatarCLIP: Zero-Shot Text-Driven Generation and Animation of 3D Avatars. *TOG 2022.*

[3] OmniObject3D: Large-Vocabulary 3D Object Dataset for Realistic Perception, Reconstruction and Generation. *CVPR 2023.*

[4] Text2Light: Zero-Shot Text-Driven HDR Panorama Generation. *TOG 2022.*

[5] Otter: A Multi-Modal Model with In-Context Instruction Tuning. *ArXiv 2023.*

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

ACM ISBN 979-8-4007-1193-0/24/10.
https://doi.org/10.1145/3688866.3689384

Multimodal Video Understanding and Generation

Mike Zheng Shou
Show Lab, National University of Singapore
Singapore
mike.zheng.shou@gmail.com

ABSTRACT

Exciting progress has been made in multimodal video intelligence, including both understanding and generation, these two pillars in video. Despite being promising, several key challenges still remain. In this talk, I will introduce our attempts to address some of them. (1) For understanding, I will share All-in-one, which employs one single unified network for efficient video-language modeling, and EgoVLP, which is the first video-language pre-trained model for egocentric video. (2) For generation, I will introduce our study of efficient video diffusion models (i.e., Tune-A-Video, 4K GitHub stars). (3) Finally, I would like to discuss our recent exploration, Show-o, one single LLM that unifies multimodal understanding and generation.

CCS Concepts/ACM Classifiers

• Computing methodologies~Artificial intelligence~Computer vision~Computer vision tasks~Video summarization.

Author Keywords

Multimodal; Video Understanding; Video Generation

ACM Reference format:
Mike Zheng Shou. 2024. Multimodal Video Understanding and Generation. 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, 1 pages. https://doi.org/10.1145/3688866.3689385

BIOGRAPHY

Prof Mike Zheng Shou joined National University of Singapore as a tenure-track Assistant Professor in 2021. He was a research scientist at Facebook AI in the Bay Area. He obtained his Ph.D. degree at Columbia University, working with Prof. Shih-Fu Chang. He received the best paper finalist at CVPR 2022, the best student paper nomination at CVPR 2017, EgoVis Distinguished Paper Award 2022/23. His team won 1st place in the international challenges including ActivityNet 2017, EPIC-Kitchens 2022, Ego4D 2022 & 2023. He is on the Forbes 30 Under 30 Asia list. He is a Singapore Technologies Engineering Distinguished Professor and a Fellow of National Research Foundation Singapore. He is on the Forbes 30 Under 30 Asia list.

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

ACM ISBN 979-8-4007-1193-0/24/10.
https://doi.org/10.1145/3688866.3689385

Leveraging the Syntactic Structure of the Text Prompt to Enhance Object-Attribute Binding in Image Generation

Maria Mihaela Trusca[*]
mariamihaela.trusca@kuleuven.be
Department of Computer Science, KU Leuven
Leuven, Belgium

Wolf Nuyts[*]
wolf.nuyts@kuleuven.be
Department of Computer Science, KU Leuven
Leuven, Belgium

Jonathan Thomm
jonathan.thomm@inf.ethz.ch
Department of Computer Science, ETH Zurich
Zürich, Switzerland

Robert Hönig
robert.hoenig@inf.ethz.ch
Department of Computer Science, ETH Zurich
Zürich, Switzerland

Thomas Hofmann
thomas.hofmann@inf.ethz.ch
Department of Computer Science, ETH Zurich
Zürich, Switzerland

Tinne Tuytelaars
tinne.tuytelaars@kuleuven.be
Department of Electrical Engineering, KU Leuven
Leuven, Belgium

Marie-Francine Moens
sien.moens@kuleuven.be
Department of Computer Science, KU Leuven
Leuven, Belgium

Abstract

Current diffusion models can generate photorealistic images from text prompts but often struggle to correctly associate the attributes mentioned in the text with the appropriate objects in the image. To address this issue, we propose focused cross-attention (FCA), which controls visual attention maps using syntactic constraints from the input sentence. Additionally, the syntactic structure of the prompt aids in disentangling the multimodal CLIP embeddings commonly used in text-to-image (T2I) generation. The resulting DisCLIP embeddings and FCA can be easily integrated into state-of-the-art diffusion models without requiring additional training. We demonstrate significant improvements in T2I generation, particularly in the accurate binding of attributes to objects, across multiple datasets.

CCS Concepts

• **Computing methodologies → Computer vision**.

Keywords

Focused Cross-Attention, Disentangle CLIP, Text-to-Image Generation, Diffusion Models

ACM Reference Format:
Maria Mihaela Trusca, Wolf Nuyts, Jonathan Thomm, Robert Hönig, Thomas Hofmann, Tinne Tuytelaars, and Marie-Francine Moens. 2024. Leveraging the Syntactic Structure of the Text Prompt to Enhance Object-Attribute Binding in Image Generation. 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, 5 pages. https://doi.org/10.1145/3688866.3689118

1 Introduction

Text-to-image synthesis (T2I) refers to the process of generating visual content based on textual input, with the goal of creating realistic images that accurately match the provided descriptions. Recent advances in this field are mainly attributed to the introduction of large-scale diffusion models trained on millions of text-image pairs, such as DALL-E 2 [15], GLIDE [12], Imagen [19], and open-source models like Stable Diffusion [18]. While these models produce high-quality, photorealistic images, their performance declines when multiple objects are mentioned in the textual prompts due to incorrect attribute-to-object binding [3, 15, 17, 19]. These models often associate objects with their most common attributes; for example, given the prompt "a golden car and a red watch," the model might generate a golden watch and a red car. Additionally, they can spread an attribute's influence across multiple objects (attribute leakage). For instance, the prompt "a golden ingot and fish" might result in a goldfish and a golden ingot [17].

To address the issue of incorrect binding between objects and their attributes, we propose two components that leverage the syntactic structure of text prompts and integrate them into diffusion-based models without requiring additional training. The term "training-free" refers to the use of pre-trained diffusion models on large-scale data, as commonly done in the literature. The first component, focused cross-attention (FCA), constrains the visual attention maps

[*]Both authors contributed equally to this research.

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

ACM ISBN 979-8-4007-1193-0/24/10
https://doi.org/10.1145/3688866.3689118

using the syntactic structure found in the input sentence. FCA ensures that the attention given to attributes is restricted to the same spatial locations as their corresponding objects. The second component uses the syntactic structure to disentangle the multimodal prompt embeddings commonly used in T2I generation. We introduce a novel encoding called *disentangled CLIP* (DisCLIP), which relies on a syntax parser to generate a constituency tree from the sentence, mitigating the entanglement issues observed with traditional CLIP encodings. Both FCA and DisCLIP lead to improved attribute binding and a reduction in attribute leakage. An additional benefit is their easy integration into any diffusion-based T2I model.

2 Related Work

The introduction of diffusion models [12], combined with classifier-free guidance [8], has led to significant improvements in image quality. However, Rassin et al. [17] highlight that T2I models still face issues with concept leakage and homonym duplication. Additionally, Petsiuk et al. [13] and Binyamin et al. [2] show that these models perform poorly on sentences containing multiple objects, attributes, and relationships. To address these problems, many models have improved spatial control of image generation by leveraging spatial constraints in the form of scene layouts to guide the diffusion process. This guidance is typically achieved by using additional resources that detect objects and their bounding boxes or segments in the image and by exploiting the object label and region associations of the attention maps (e.g., [1, 4, 5, 7, 9, 10, 21, 23, 24, 26]), or by using sketches, as done by ControlNet [28]. In this work, we do not rely on such additional resources but instead use the syntactic structure of the text prompt to provide guidance.

Feng et al. [6] adapt the Stable Diffusion backbone to attend to multiple encodings representing the syntactic constituents of the text prompt. Similar to this work, which we use as a baseline, we leverage the syntax of the text prompt but explicitly exploit syntactic dependencies to bind attributes to objects, leading to better T2I generation. Attend-and-Excite [3] improves the cross-attention between objects mentioned in the text prompt and the image embeddings, demonstrating that their method is particularly suited for generating multiple objects. Zhang et al. [27] address the generation of multiple objects by learning their masked regions in the image. Using a set of loss functions, a diffusion-based model gradually learns to allocate the objects and their attributes to the designated masked regions and to prevent overlapping over the regions of other objects and the background. SynGen [16] syntactically analyzes the prompt and uses this information in appropriate loss functions to enhance the similarity between the attention maps of objects and their attributes while increasing the distance between these attention maps and those of other words in the prompt. We show that the proposed FCA and DisCLIP encoding can be seamlessly integrated into state-of-the-art T2I generation baselines, including Attend-and-Excite and SynGen, and improve their results.

3 Preliminaries

Cross-Attention in Diffusion Models. The diffusion models [12, 18, 19] are defined based on U-Nets that use cross-attention layers to condition a denoising network ϵ_θ on a text prompt y. A common implementation of this cross-attention uses query (here encoded image), key and value (here encoded text) attention of Vaswani et al. [22] to calculate the cross-attention maps $A^l_t \in \mathbb{R}^{h\times w,n}$ for each layer l and timestep t of the denoising process:

$$Q^l_t = x^l_t W^l_Q,\ K^l = yW^l_K,\ V^l = yW^l_V \tag{1}$$

$$A^l_t = softmax(\frac{Q^l_t (K^l)^{\mathrm{T}}}{\sqrt{d}}),\ f^l_t = A^l_t V^l \tag{2}$$

where W^l_Q represents a linear layer transforming x^l_t into the queries $Q^l_t \in \mathbb{R}^{h\times w,d}$, where d denotes the feature dimension. Similarly, W^l_K and W^l_V transform y into keys $K^l \in \mathbb{R}^{n,d}$ and values $V^l \in \mathbb{R}^{n,d}$. $f^l_t \in \mathbb{R}^{h\times w,d}$ represents the output features of the cross-attention layer.

4 Methods to Improve Object-Attribute Binding in T2I Generation

We propose two training-free methods to enhance the text conditioning of diffusion models. The first method, called focused cross-attention (FCA), utilizes a syntactic parse of the text prompt to confine the attention of an attribute to the regions where the corresponding object is active. This approach integrates seamlessly with diffusion models that rely on large language encoders trained solely on text. The second method introduces a new disentangled CLIP representation (DisCLIP), addressing the attribute binding issues found in standard CLIP embeddings [15]. DisCLIP also incorporates a syntactic parse of the text prompt.

4.1 Focused Cross-Attention (FCA)

To improve the binding of attributes to the correct objects, we restrict the attention of attributes to regions where their corresponding object has attention as well. Attribute dependencies are obtained from a dependency parse of the sentence and implemented in a binary matrix $D \in \{0,1\}^{n\times n}$, representing for each token of y the token on which it is dependent. [1] Diffusion with FCA operates using two denoising model traversals, as formalized in Algorithm 1. In the first traversal, standard cross-attention A^l_t is used, from which the average attention maps A^* are obtained by averaging A^l_t over each layer l and timestep t. From these attention maps, we obtain the focus mask $F_{\text{mask}} \in \{-\infty, 0\}^{h\times w,n}$ and calculate FCA as follows:

$$F_{\text{mask}} = \delta\big(A^* D^{\mathrm{T}}\big),\ \text{with} \tag{3}$$

$$\delta\big(b_{ij}\big) = -\infty \text{ if } \frac{b_{ij} - \min\limits_p\big(b_{pj}\big)}{\max\limits_p\big(b_{pj}\big) - \min\limits_p\big(b_{pj}\big)} < s,\ \text{else } 0 \tag{4}$$

$$\text{FCA}\big(Q, K, V, F_{\text{mask}}\big) = \text{softmax}\big(\frac{F_{\text{mask}} + QK^{\mathrm{T}}}{\sqrt{d}}\big)\ V \tag{5}$$

where the threshold s is a hyperparameter and δ is a function that operates on each cell b_{ij} of $A^* D^{\mathrm{T}}$. $F_{\text{mask}} = -\infty$ for the attributes' cross-attention map regions where its corresponding object has a normalized attention map value less than s. By replacing the cross-attention of $\epsilon_\theta(x_t, y, t)$ with FCA (Equation 5), we obtain $\epsilon_{\theta,FCA}(x_t, y, t, F_{\text{mask}})$. $\epsilon_{\theta,FCA}$ is then used to obtain the output image I* in a second model traversal with FCA. Dimensions $w \times h$ are

[1] The dependency matrix can implement complex relationships involving multiple objects and their respective attributes.

Algorithm 1 Diffusion with FCA

```
Input: sentence encoding $y$ attribute dependencies $D$
$x_T \leftarrow N(0, I)$
for $t \leftarrow T \ldots 1$ do
    $z_{t-1}, \{A_t^l\} \leftarrow \epsilon_\theta(x_t, y, t)$
    $x_{t-1} \leftarrow sample(x_t, z_{t-1})$
end for
$A^* \leftarrow \overline{A_t^l}$
$F_{\text{mask}} \leftarrow \delta(A^* D^{\mathrm{T}})$
$x_T^* \leftarrow x_T$
for $t \leftarrow T \ldots 1$ do
    $z_{t-1}^* \leftarrow \epsilon_{\theta,FCA}(x_t^*, y, t, F_{\text{mask}})$
    $x_{t-1}^* \leftarrow sample(x_t^*, z_{t-1}^*)$
end for
Output: image $\mathrm{I}^* \leftarrow x_0^*$
```

not the same for different layers l in ϵ_θ. We use cubic interpolation to the largest layer size to average A_t^l over layers of different sizes. Max pooling is used to project F_{mask} back to the correct layer size of l.

4.2 Disentangled CLIP Encoding (DisCLIP)

A CLIP encoding of a sentence includes embeddings of each word, concatenated with a sentence embedding and padding embeddings [14]. T2I models based on CLIP encodings often struggle with image-text alignment [19]. We propose a novel training-free variation of CLIP, called DisCLIP. DisCLIP utilizes a syntactic parser to derive a hierarchical representation of the text prompt in the form of a constituency tree. By replacing noun phrases in the higher layers of the tree with their head nouns, we create an abstracted constituency tree.This tree encodes compositional information, including explicit object-attribute bindings. The tree is then used to disentangle the CLIP representation of the text prompt. We independently encode the entire prompt or sentence, and each constituent of the tree with CLIP (removing padding embeddings), and concatenate the resulting embeddings. When used with FCA, an extra row and column are added to D for each additional constituent embedding, indicating a dependency between the added constituent and the nouns present within it. The results below demonstrate that DisCLIP mitigates the problem of object-attribute binding.

5 Experimental Set-up

5.1 Datasets and Metrics

We evaluate the object-attribute binding of the models on Concept Conjunction 500 (CC-500) [6] and the Attend-and-Excite (AE-276) [3] datasets. Additionally, we report results on a novel dataset called Difficult Adversarial Attributes (DAA-200), specially defined for evaluating object-attribute binding in T2I generation. DAA-200 uses the image-graph pairs of Visual Genome to obtain 100 quadruplets of the form {attribute 1, object 1, attribute 2, object 2}. These quadruplets can be represented in a simple graph with the two objects as nodes and one attribute for each node. From each graph, an adversarial graph is generated by swapping the attributes of both objects. For each of the 200 graphs a sentence of the form $\langle attribute1\rangle\langle object1\rangle and\langle attribute2\rangle\langle object2\rangle$." is generated. To ensure that we have difficult adversarial examples, for DAA-200 we picked examples from Visual Genome where both objects occur multiple times in Visual Genome with each of the attributes.

We use two human evaluations to assess the image fidelity and image-text alignment. First, we ask annotators to compare two generated images and indicate which image demonstrates better image-text alignment and image fidelity. Second, following Feng et al. [6], we ask annotators whether the two objects of the CC-500 samples are present in the generated images and whether they are in the correct color. We also ask whether a part of the object is in the color of the other object to assess how many attributes are leaked to the wrong objects. The human evaluation was executed with the use of the Amazon Mechanical Turk and Clickworker platforms. Results are shown in Tables 1-2.

5.2 Models and Hyperparameters

We use open source T2I diffusion models and expand these with the FCA component and DisCLIP encoding of the text prompt. We do not show results for the method introduced in [27] as the code of the method is not yet available.
Stable Diffusion SD is trained on a filtered version of the LAION-5B [20] dataset, uses the latent diffusion architecture of Rombach et al. [18] and uses a frozen CLIP [14] model as the text encoder.
Attend-and Excite AE refers to the original Attend-and-Excite model [3]. It builds on Stable Diffusion and is designed to improve the generating of multiple objects mentioned in the text prompt by focusing the attention on nouns appearing in the text prompt.
Versatile Diffusion VD extends an existing single flow diffusion pipeline into a multitask multimodal network that handles T2I, image-to-text and image-variation generation [25].
SynGen [16] relies on loss functions to align objects with their attributes.
Structure Diffusion adapts the Stable Diffusion model to attend to multiple encodings of syntax constituents of the text prompt [6]. The above models are used as baselines.
$\mathbf{SD_{FCA+DisCLIP}}$, $\mathbf{AE_{FCA+DisCLIP}}$, $\mathbf{SynGen_{FCA+DisCLIP}}$ and $\mathbf{VD_{FCA+DisCLIP}}$ integrate FCA and DisCLIP into Stable Diffusion, Attend-and-Excite, SynGen and Versatile Diffusion, respectively.

All comparisons use the same seeds for each model with 50 diffusion steps and a guidance scale of 7.5. Dependency and constituency parses are obtained with the LAL-parser of Mrini et al. [11]. The hyperparameter s used to implement FCA is set to 0.6. The value was selected by measuring the classification accuracy in % on the ground truth images of DAA-200. When quantitatively evaluating the T2I generation, for DAA-200, we generate 10 images per prompt; for CC-500 we follow Feng et al. [6] and generate 3 images per prompt; for AE-276 we generate three images per prompt.

6 Results and Discussion

In this section, we discuss the results obtained on datasets that challenge object-attribute binding, which are DAA-200, CC-500 and AE-276 datasets. Table 1 presents the results of the human evaluation. Observe that for DAA-200, CC-500 and AE-276 our methods outperform the baselines considering image-text alignment and all baselines but one with regard to image fidelity. The largest increase is seen on CC-500 where our methods outperform other models by 4-25 percentage points on alignment and 8-15 percentage points on fidelity. We hypothesize that T2I generation struggles when it is not

Table 1: Percentage of cases in which our FCA and DisCLIP modules generate better (win) or worse (lose) alignment and T2I fidelity than their baselines. Structure Diffusion is an adaptation of the Stable Diffusion (*SD*) for object-attribute binding. Therefore Stable Diffusion enhanced with our modules, FCA and DisCLIP (SD$_{\text{FCA+DisCLIP}}$) is compared only with the Structure Diffusion.

	ours	Alignment		Fidelity	
Benchmark	**v.s.**	Win ↑	Lose ↓	Win ↑	Lose ↓
DAA-200	StructureDiffusion	**36.1**	34.0	**37.0**	32.9
	Attend-and-Excite	**32.5**	28.5	37.0	**38**
	Versatile Diffusion	**34.2**	31.3	**37.3**	33.8
	SynGen	**33.2**	30.3	**36.2**	32.7
CC-500	StructureDiffusion	**32.4**	28.2	**43.3**	28.4
	Attend-and-Excite	**49.7**	24.5	**45.8**	31.8
	Versatile Diffusion	**28.1**	27.7	**37.5**	33.1
	SynGen	**36.5**	34.4	**39.7**	34.7
AE-276	StructureDiffusion	**36.2**	27.8	**45.6**	33.3
	Attend-and-Excite	**36.3**	27.1	**39.8**	36.9
	Versatile Diffusion	**29.3**	27.3	29.2	**30.7**
	SynGen	31.3	**34.2**	34.6	**35.1**

Table 2: Results of the human evaluation obtained on CC-500. We show how often (in %) each model is able to correctly (with the correct color) generate at least one object / the two objects of the CC-500 captions. Leakage displays how often (in %) an object is at least partially generated with the color of the wrong object.

Methods	Two objects ↑	Atleast one object ↑	Leakage ↓
Stable Diffusion	20.7	76.9	64.9
StructureDiffusion	21.2	77.2	63.9
SD$_{\text{DisCLIP+FCA}}$ (ours)	**22.2**	**80.8**	**56.8**
Attend-and-Excite AE	46.8	88.4	65.4
AE$_{\text{FCA+DisCLIP}}$ (ours)	**60.2**	**94.3**	**64.5**
Versatile Diffusion VD	23.4	72.8	77.5
VD$_{\text{FCA+DisCLIP}}$ (ours)	**25.6**	**76.3**	**69.7**
SynGen	45.3	90.3	32.1
SynGen$_{\text{FCA+DisCLIP}}$ (ours)	**47.2**	**91.2**	**27.4**

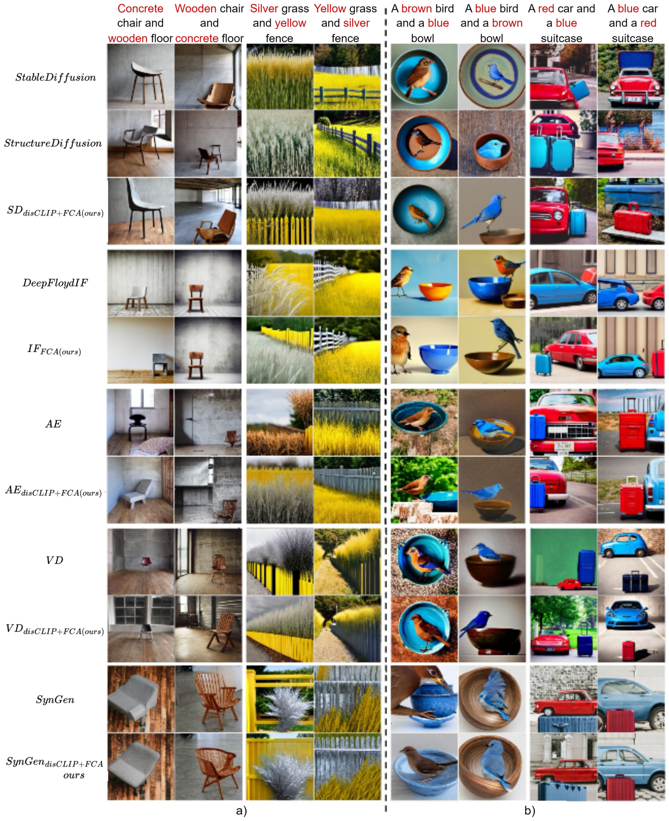

Concrete chair and wooden floor
Wooden chair and concrete floor
Silver grass and yellow fence
Yellow grass and silver fence
A brown bird and a blue bowl
A blue bird and a brown bowl
A red car and a blue suitcase
A blue car and a red suitcase
StableDiffusion
StructureDiffusion
SD disCLIP+FCA(ours)
DeepFloydIF
IF FCA(ours)
AE
AE disCLIP+FCA(ours)
VD
VD disCLIP+FCA(ours)
SynGen
SynGen disCLIP+FCA ours
a)
b)

Figure 1: Qualitative results that show that the FCA and DisCLIP enhanced models improve attribute binding and decrease attribute leakage in images from (a) DAA-200 and (b) CC-500.

straightforward which attribute belongs to which object. Because color attributes often co-occur with many objects, the captions of CC-500 are especially difficult for T2I models (as they contain only color attributes). This leads to a larger improvement for our models that explicitly bind attributes to certain regions, as can be seen in Figure 1b. DAA-200 and AE-267, on the other hand, contain diverse categories of attributes. Base models are good at generating the most expected attribute binding. Unlike our models, they perform poorly when attributes are switched. An example is shown in Figure 1a where all models perform well on the prompt "yellow grass and silver fence" but only our models perform well on the prompt "Silver grass and yellow fence". Although SynGen already achieves accurate object-attribute alignments, when augmenting this model with FCA and DisCLIP, the image quality is enhanced the leakage is decreased. Table 2 successfully evaluates object-attribute binding of the proposed methods by conducting a human evaluation that checks whether each object of CC-500 is generated with the correct color. We perform this analysis on CC-500 as the attributes are only colors that are associated with a wide range of objects. SD$_{\text{DisCLIP+FCA}}$ and IF$_{\text{FCA}}$ decrease leakage by 8 and 2.5 percentage points, respectively.

7 Conclusion

We have proposed training-free methods to emphasize the importance of integrating linguistic syntactic structures in T2I generation. We demonstrated their easy and successful integration in state-of-the-art T2I diffusion models leading to an improved object-attribute binding and to a decrease in attribute leakage in the generated image.

Acknowledgments

This project was funded by the European Research Council (ERC) Advanced Grant CALCULUS (grant agreement No. 788506).

References

[1] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. 2023. SpaText: Spatio-Textual Representation for Controllable Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18370–18380.

[2] Lital Binyamin, Yoad Tewel, Hilit Segev, Eran Hirsch, Royi Rassin, and Gal Chechik. 2024. Make It Count: Text-to-Image Generation with an Accurate Number of Objects. *CoRR* abs/2406.10210 (2024).

[3] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models. *ACM Trans. Graph.* 42, 4 (2023), 148:1–148:10. https://doi.org/10.1145/3592116

[4] Hongyu Chen, Yiqi Gao, Min Zhou, Peng Wang, Xubin Li, Tiezheng Ge, and Bo Zheng. 2024. Enhancing Prompt Following with Visual Control Through Training-Free Mask-Guided Diffusion. *CoRR* abs/2404.14768 (2024).

[5] Minghao Chen, Iro Laina, and Andrea Vedaldi. 2024. Training-Free Layout Control with Cross-Attention Guidance. In *Proceedings of WACV*.

[6] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun R. Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2023. Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis. In *11th International Conference on Learning Representations, ICLR 2023*. https://openreview.net/pdf?id=PUIqjT4rzq7

[7] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Prompt-to-Prompt Image Editing with Cross-Attention Control. In *11th International Conference on Learning Representations, ICLR 2023*. https://openreview.net/pdf?id=_CDixzkzeyb

[8] Jonathan Ho and Tim Salimans. 2022. Classifier-Free Diffusion Guidance. arXiv:2207.12598 [cs.LG]

[9] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023. GLIGEN: Open-Set Grounded Text-to-Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 22511–22521.

[10] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. 2022. Compositional Visual Generation with Composable Diffusion Models. In *17th European Conference on Computer Vision, ECCV 2022*. Springer, 423–439. https://doi.org/10.1007/978-3-031-19790-1_26

[11] Khalil Mrini, Franck Dernoncourt, Quan Hung Tran, Trung Bui, Walter Chang, and Ndapa Nakashole. 2020. Rethinking Self-Attention: Towards Interpretability in Neural Parsing. In *Findings of the Association for Computational Linguistics, EMNLP 2020*. Association for Computational Linguistics, 731–742. https://doi.org/10.18653/v1/2020.findings-emnlp.65

[12] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *39th International Conference on Machine Learning, ICML 2022*. PMLR, 16784–16804. https://proceedings.mlr.press/v162/nichol22a.html

[13] Vitali Petsiuk, Alexander E. Siemenn, Saisamrit Surbehera, Zad Chin, Keith Tyser, Gregory Hunter, Arvind Raghavan, Yann Hicke, Bryan A. Plummer, Ori Kerret, Tonio Buonassisi, Kate Saenko, Armando Solar-Lezama, and Iddo Drori. 2022. Human Evaluation of Text-to-Image Models on a Multi-Task Benchmark. arXiv:2211.12112 [cs.CV]

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *38th International Conference on Machine Learning, ICML 2021*. PMLR, 8748–8763. http://proceedings.mlr.press/v139/radford21a.html

[15] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125 [cs.CV]

[16] Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. 2023. Linguistic Binding in Diffusion Models: Enhancing Attribute Correspondence through Attention Map Alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=AOKU4nRw1W

[17] Royi Rassin, Shauli Ravfogel, and Yoav Goldberg. 2022. DALLE-2 is Seeing Double: Flaws in Word-to-Concept Mapping in Text2Image Models. In *5th BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2022*. 335–345. https://aclanthology.org/2022.blackboxnlp-1.28

[18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*. 10674–10685. https://doi.org/10.1109/CVPR52688.2022.01042

[19] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *36th Conference on Neural Information Processing Systems, NeurIPS 2022*. http://papers.nips.cc/paper_files/paper/2022/hash/ec795aeadae0b7d230fa35cbaf04c041-Abstract-Conference.html

[20] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *36th Conference on Neural Information Processing Systems, NeurIPS 2022*. http://papers.nips.cc/paper_files/paper/2022/hash/a1859debfb3b59d094f3504d5ebb6c25-Abstract-Datasets_and_Benchmarks.html

[21] Ashkan Taghipour, Morteza Ghahremani, Mohammed Bennamoun, Aref Miri Rekavandi, Hamid Laga, and Farid Boussaïd. 2024. Box It to Bind It: Unified Layout Control and Attribute Binding in T2I Diffusion Models. *CoRR* abs/2402.17910 (2024).

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *30th Conference on Neural Information Processing Systems, NeurIPS 2016*. 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[23] Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. 2024. InstanceDiffusion: Instance-level Control for Image Generation. arXiv:2402.03290 [cs.CV]

[24] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. 2023. BoxDiff: Text-to-Image Synthesis with Training-Free Box-Constrained Diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 7452–7461.

[25] Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. 2023. Versatile Diffusion: Text, Images and Variations All in One Diffusion Model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 7754–7765.

[26] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. ReCo: Region-Controlled Text-to-Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14246–14255.

[27] Jiewei Zhang, Song Guo, Peiran Dong, Jie Zhang, Ziming Liu, Yue Yu, and Xiao-Ming Wu. 2024. Easing Concept Bleeding in Diffusion via Entity Localization and Anchoring. In *Forty-first International Conference on Machine Learning*.

[28] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 3836–3847.

Geo-LLaVA: A Large Multi-Modal Model for Solving Geometry Math Problems with Meta In-Context Learning

Shihao Xu*
xush0019@ntu.edu.sg
Huawei Singapore Research Center
Singapore, Singapore

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

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

Abstract

Geometry mathematics problems pose significant challenges for large language models (LLMs) because they involve visual elements and spatial reasoning. Current methods primarily rely on symbolic character awareness to address these problems. Considering geometry problem solving is a relatively nascent field with limited suitable datasets and currently almost no work on solid geometry problem solving, we collect a geometry question-answer dataset by sourcing geometric data from Chinese high school education websites, referred to as GeoMath. It contains solid geometry questions and answers with accurate reasoning steps as compensation for existing plane geometry datasets. Additionally, we propose a Large Multi-modal Model (LMM) framework named Geo-LLaVA, which incorporates retrieval augmentation with supervised fine-tuning (SFT) in the training stage, called meta-training, and employs in-context learning (ICL) during inference to improve performance. Our fine-tuned model with ICL attains the state-of-the-art performance of 65.25% and 42.36% on selected questions of the GeoQA dataset and GeoMath dataset respectively with proper inference steps. Notably, our model initially endows the ability to solve solid geometry problems and supports the generation of reasonable solid geometry picture descriptions and problem-solving steps. Our research sets the stage for further exploration of LLMs in multi-modal math problem-solving, particularly in geometry math problems.

CCS Concepts

• **Computing methodologies** → **Knowledge representation and reasoning**; **Natural language processing**; **Computer vision representations**.

Keywords

Large Multimodal Model, Geometry Problem Solving, RAG, In-Context Learning

ACM Reference Format:
Shihao Xu, Yiyang Luo, and Wei Shi. 2024. Geo-LLaVA: A Large Multi-Modal Model for Solving Geometry Math Problems with Meta In-Context Learning. 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, 5 pages. https://doi.org/10.1145/3688866.3689124

1 Introduction

There has been an increased interest in using deep learning models, especially LMMs, for addressing computer vision challenges recently. Aligning image adapters with large language models has achieved remarkable success in image captioning and visual question answering (VQA), highlighting their powerful reasoning and visual understanding capabilities [8]. Despite their accomplishments in these areas, there has been limited investigation into utilizing LLMs for more complex multi-modal math problems, particularly geometry-related ones.

The debate between symbolic and probabilistic approaches in mathematical reasoning persists. Traditionally, solving geometry problems involves analyzing diagrams and texts, converting them into logical expressions using formal symbolic language, and applying predefined geometry theorems to find solutions [11, 33]. Alternatively, geometry problem-solving can be viewed as text generation with multi-modal input, which is more generalized and applicable to a broader range of mathematical problems, including trigonometry and vector graphics. By examining probabilistic approaches, we can better understand their strengths and limitations, leading to more effective strategies for solving geometry problems.

Current Large Multi-modal Models (LMMs) have shown promising capabilities in visual understanding and question-answering tasks, as demonstrated by models such as BLIP-2 [20], LLaVA [22], Flamingo [3], MiniGPT4 [34], and InstructBLIP [12]. However, these models still lack a deep comprehension of geometry images, which is crucial for solving geometry problems. Additionally, existing small language models often lack the mathematical reasoning abilities to solve complex math problems effectively.

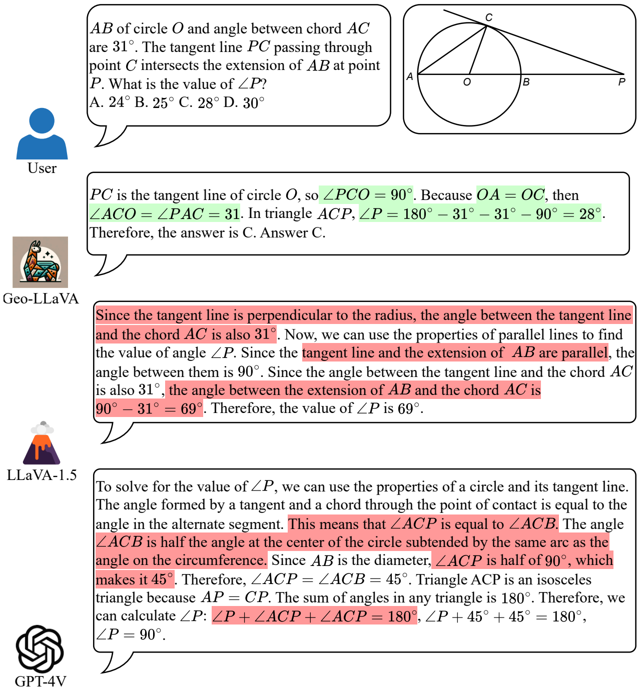
In this paper, we propose a solution to the challenges of multi-modal math problem-solving by introducing an LLM framework called Geo-LLaVA. The main contributions of this paper include:

- We form a geometry question-answer dataset, GeoMath, with reasoning steps from Chinese high school education websites and expand the dataset by collecting more data from existing datasets such as GeoQA+ [11], Unigeo [10], and PGPS9K [33], and creating reasoning steps for them.
- We employ a new LLM framework named Geo-LLaVA with around 13 billion parameters. It can effectively generate reasoning steps and answers, which is the first model exploring both plane and solid geometry problems.

*Both authors contributed equally to this research.

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

ACM ISBN 979-8-4007-1193-0/24/10
https://doi.org/10.1145/3688866.3689124



AB of circle O and angle between chord AC are 31°. The tangent line PC passing through point C intersects the extension of AB at point P. What is the value of $\angle P$?
A. 24° B. 25° C. 28° D. 30°

PC is the tangent line of circle O, so $\angle PCO = 90^\circ$. Because $OA = OC$, then $\angle ACO = \angle PAC = 31$. In triangle ACP, $\angle P = 180^\circ - 31^\circ - 31^\circ - 90^\circ = 28^\circ$. Therefore, the answer is C. Answer C.

Since the tangent line is perpendicular to the radius, the angle between the tangent line and the chord AC is also 31°. Now, we can use the properties of parallel lines to find the value of angle $\angle P$. Since the tangent line and the extension of AB are parallel, the angle between them is 90°. Since the angle between the tangent line and the chord AC is also 31°, the angle between the extension of AB and the chord AC is $90^\circ - 31^\circ = 69^\circ$. Therefore, the value of $\angle P$ is 69°.

To solve for the value of $\angle P$, we can use the properties of a circle and its tangent line. The angle formed by a tangent and a chord through the point of contact is equal to the angle in the alternate segment. This means that $\angle ACP$ is equal to $\angle ACB$. The angle $\angle ACB$ is half the angle at the center of the circle subtended by the same arc as the angle on the circumference. Since AB is the diameter, $\angle ACP$ is half of 90°, which makes it 45°. Therefore, $\angle ACP = \angle ACB = 45^\circ$. Triangle ACP is an isosceles triangle because $AP = CP$. The sum of angles in any triangle is 180°. Therefore, we can calculate $\angle P$: $\angle P + \angle ACP + \angle ACP = 180^\circ$, $\angle P + 45^\circ + 45^\circ = 180^\circ$, $\angle P = 90^\circ$.

Figure 1: Example of geometry question answering. LLaVA-1.5 often provides a vague image description, omitting many details and leading to wrong answers. GPT-4V struggles to understand geometry graphs, resulting in incorrect answers. In contrast, our Geo-LLaVA offers concise and accurate solutions to geometry problems.

- As far as we know, this is the first multimodal meta-training [24] method to enhance geometry problem-solving ability, which shows the state-of-the-art performance on multiple geometry datasets, including GeoQA+ and GeoMath.

2 Related Work

Meta-Training Approaches. The general issue of meta-training [28], which encompasses few-shot learning, has been studied for numerous years. Recently, meta-training has emerged as a significant technique in machine learning. Using prior knowledge, it aims to create models that can swiftly adapt to new tasks with minimal data. meta-training is particularly pertinent in geometry problem-solving, where the diversity of problems can vary greatly. The most recent meta-training models, such as MAML [15], Hypernetworks [18], and [27], have shown promise in rapidly adapting to new tasks with few-shot learning capabilities.

Multi-Modal Large Language Model. Concurrently, the success of LLMs has inspired investigations into vision-language interaction, resulting in the development of multi-modal large language models (MLLMs) [1, 7, 12, 20, 22, 32]. These models have demonstrated remarkable abilities in generating detailed descriptions and engaging in dialogue based on visual inputs. Nonetheless, we observe that even the most advanced MLLMs struggle with resolving geometric problems using diagrams and figures.

Geometry Problem Solving by LLMs. Recently, math-specific LLMs such as Llemma [6] and Mathcoder [31] have shown significant capabilities in text-only mathematical reasoning tasks and are competitive with general large language models like GPT-4 [2] and PaLM-2 [4] on a much smaller scale. Notably, AlphaGeometry [30] has exhibited impressive performance in solving challenging geometry problems, though it cannot process images and must rely on text descriptions. Current math-specific multimodal models, such as G-llava [16], UniMath [21], and UniChart [23], are primarily focused on plane geometry or chart-based problems and still lag behind general multimodal models such as GPT-4V [1] in benchmark testing. Yet, no works have utilized RAG or meta-training techniques on geometric problem-solving.

3 Method

The Geo-LLaVA model, illustrated in Figure 1, consists of a retrieval network and an LMM backbone. The retrieval network’s role is to fetch similar questions and their solutions as in-context samples during the training and inference phases. In the following section, we offer a comprehensive overview of the design process for our retrieval network. Next, we delve into our method of fine-tuning the model using image captioning, question-answering, and geometry math-solving datasets through Meta in-context learning. Lastly, we will describe how we have integrated a multi-modal chain of thoughts (CoT) during the inference phase to boost the model’s performance even further.

3.1 Retrieval Network

In this study, similar to CLIP [26], we implemented a dual-tower network framework for retrieval tasks. Specifically, the pre-trained ViT-L-14 [14] and Bert [13] (Bert-base-uncased) models were applied as the image and language encoders, respectively. We integrated two adapter layers into each encoder to ensure compatibility between these encoders. These adapter layers comprise three linear layers with ReLU activation functions, designed to harmonize the embedding dimensions of both encoders.

The Bert model [13] was chosen as the language encoder due to its strong performance in natural language processing tasks. Importantly, we transformed math-specific tokens absent in the BERT pre-trained vocabulary into words (e.g., △ → triangle, ⊥→ perpendicular to) during the preprocessing stage. We believe the BERT model can effectively grasp the meaning of inputs related to geometric problems. Similarly, the ViT-L-14 [14] model was selected as the image encoder for its outstanding performance in image recognition tasks. This model employs a transformer-based architecture, which is particularly adept at processing visual information. However, the pre-trained ViT model may not generalize well to geometric math images. As a result, we retrained the parameters of the two adapter layers and the ViT model from scratch using question-image pairs with the InfoNCE loss, a contrastive learning technique recognized for its efficacy in training neural networks for retrieval tasks. This loss function encourages the model to learn meaningful representations of the input question and the corresponding image, promoting accurate information retrieval.

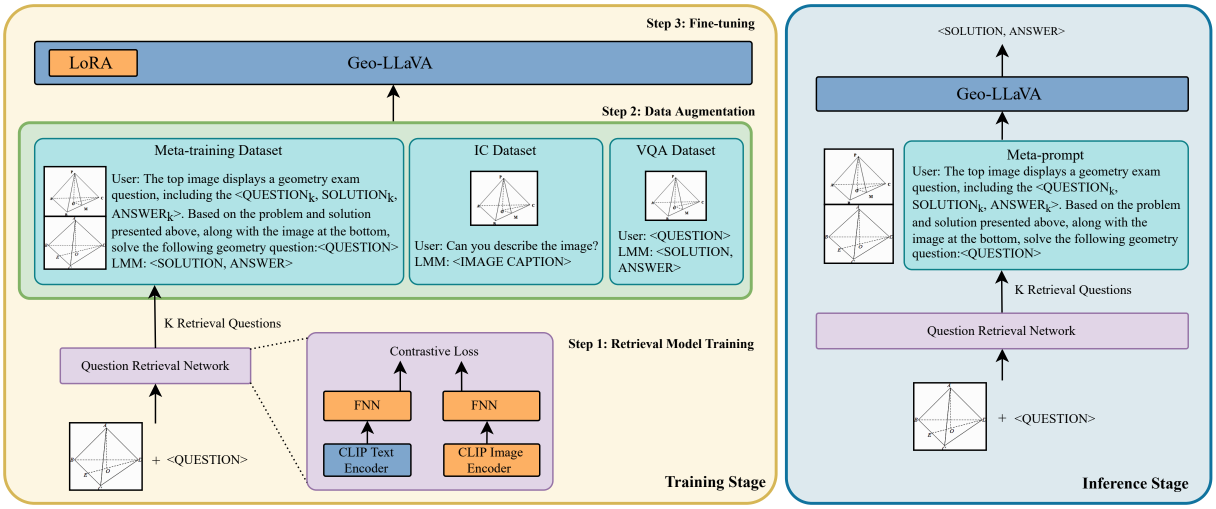

Step 3: Fine-tuning
LoRA
Geo-LLaVA
Step 2: Data Augmentation
Meta-training Dataset
User: The top image displays a geometry exam question, including the <QUESTION$_k$, SOLUTION$_k$, ANSWER$_k$>. Based on the problem and solution presented above, along with the image at the bottom, solve the following geometry question:<QUESTION>
LMM: <SOLUTION, ANSWER>
IC Dataset
User: Can you describe the image?
LMM: <IMAGE CAPTION>
VQA Dataset
User: <QUESTION>
LMM: <SOLUTION, ANSWER>
K Retrieval Questions
Question Retrieval Network
+ <QUESTION>
Contrastive Loss
FNN
FNN
CLIP Text Encoder
CLIP Image Encoder
Step 1: Retrieval Model Training
Training Stage
<SOLUTION, ANSWER>
Geo-LLaVA
Meta-prompt
User: The top image displays a geometry exam question, including the <QUESTION$_k$, SOLUTION$_k$, ANSWER$_k$>. Based on the problem and solution presented above, along with the image at the bottom, solve the following geometry question:<QUESTION>
K Retrieval Questions
Question Retrieval Network
+ <QUESTION>
Inference Stage

Figure 2: The overview of Geo-LLaVA model architecture. The training process includes three steps: training the retrieval model, augmenting data with the meta-training dataset, and fine-tuning the Language Model with the LoRA module. The question retrieval network is trained using inputs from a CLIP text encoder and a CLIP image encoder, supervised by a contrastive loss. FNNs represent feedforward networks. During the fine-tuning step, a LoRA [19] is introduced for efficient fine-tuning and improved results.

3.2 LMM backbone

This study utilized the pre-trained LLaVA1.5-13B [22] as the LMM backbone. This model leverages the renowned LLaMA-2 [29] for advanced language processing tasks and incorporates the CLIP [25] visual encoder ViT-L/14 [14] for sophisticated visual comprehension. The integration involves a Multi-Layer Perceptron (MLP) based vision-language connector that aligns the outputs of the vision encoder with the language model. This alignment is crucial as it significantly enhances the model to handle and understand multi-modal data effectively.

3.3 Datasets for Multi-Modal Geometric Concepts and Reasoning

To address the limitations of existing models in understanding and reasoning about multi-modal geometric concepts, we developed three specialized geometry datasets:

- GeoMath-IC (Image-Context): This dataset includes images paired with simple yet comprehensive descriptions and aims to bridge the gap between visual inputs and textual descriptions in geometric concepts.
- GeoMath-QA (Question-Answer): This dataset focuses on providing questions with detailed reasoning steps. The questions are designed to challenge the model's understanding and reasoning capabilities regarding geometric concepts.
- GeoMath-Meta: This dataset includes a range of geometry problems that require the model to generalize from its learning on GeoMath-IC and GeoMath-QA.

These datasets were subsequently used to fine-tune the model using Low-Rank Adaptation (LoRA) [19]. Detailed information about the dataset creation process and examples can be found in Section 4.2.1.

We adopted the input format from the original Flamingo model, which efficiently integrates visual and linguistic elements to enhance their synergistic interaction. This structured input format involves specific templates and prompts that guide the model in understanding the context and relationships between the visual and textual components.

3.4 Enhancing In-Context Learning

To further improve the model's in-context learning capabilities, especially for smaller models, we explored using meta-training [24]. This approach enhances the performance of the model by providing relevant context during the fine-tuning stage. The procedure involves the following steps: 1) Contextual Retrieval: For each input sample, we retrieved the K most similar samples from the training data, ensuring that the input sample itself is excluded, where K is set to 1 in this paper. This retrieval is based on semantic similarity metrics, ensuring the context is highly relevant. 2) Concatenation and Fine-Tuning: The retrieved texts and corresponding images are concatenated with the input sample. This concatenated form is used to fine-tune the LMM, aligning it better with the reasoning required for tasks. Since the pre-trained LLaVA only supports single-image input, we vertically merge K images into a single image.

The detailed methodologies and the careful orchestration of these components can be referenced throughout the underlying sections of the study, particularly in Section 4.2.1 and Figure 2.

4 Experiment

4.1 Dataset

In this study, we collected about 10K solid geometry multimodal QA datasets from the 21st-century education website in China, named GeoMath dataset. All Chinese contents in these datasets are automatically translated to English using ChatGPT3.5. The detailed statistics of these three datasets are shown in Table 1. In addition, to supplement the data of solid geometry, we formed the geometry data from two existing datasets (GeoQA+ [9] and PSDK9K), which include images of plane geometry as well as questions and answers.

Table 1: The detailed statistics of four datasets. Each row represents the number of samples of specific types. The number inside the brackets represents the number of test samples.

Dataset Stat. type	GeoQA	PSDK-9K	UniGeo	GeoMath
QA-selection	12526(1509)	9986(1047)	-	4258(404)
QA-cloze	-	-	-	1423 (150)
QA-proving	-	-	9309(2899)	3474 (352)
Image-text pairs	4406	4000	-	4540 (453)
Provide solution	√	-	√	√

4.2 Setup Details

4.2.1 Data Augmentation. Paraphrasing by LLMs can generate a more diverse set of training examples and has been widely used for data augmentation. Similarly, we adopt GPT3.5 [2] for image caption and question-answering samples followed by a translation from Chinese to English and employ text rewriting to increase variety. We utilize an LLM to rephrase the input text in 5 different ways, resulting in a six times larger sample size for image caption and QA. For the GeoMath-Meta dataset, the retrieval model selects the top 5 samples with the highest similarity to construct the data for Metatraining. This ensures consistency in the quantity of the final samples and QA pairs.

4.2.2 Training settings. We select LLaVA-1.5 [22], a large multimodal model that combines the strengths of LLaMA-2 [29] and fine-tuned retrieval model [25], for our experiments. All experiments are conducted with consistent parameter settings during the LoRA fine-tuning phase. Specifically, we use a learning rate of 2×10^{-4} with a cosine learning rate scheduler and train the model for 5 epochs. The maximum token generation length is 2048. The batch size per GPU is 4, and we use gradient accumulation steps set to 4. We initially evaluate the experiments on the validation set to identify the best results, which are subsequently tested on the test set.

4.3 Experiment results

Table 2 summarizes the main results on GeoQA+ and GeoMath datasets. Three small-size LMMs (G-llava [16], OpenFlamingo [5] and LLaVA [22]) for both zero-shot and finetuning and two extremely large LMMs (Bard [17] and ChatGPT-4V [1]) are chosen as baseline models. We finetuned the model 5 times to calculate the mean and standard deviation of the evaluation metric.

The experiment results, summarized in Table 2, demonstrate the performance of various models across the GeoQA+ and GeoMath datasets, alongside an ablation study exploring the incremental application of techniques in our model. The proposed Geo-LLaVA-13B model showed significant improvements compared to the GPT-4V and Bard through the sequential addition of the GeoMath-IC, GeoMath-QA, and GeoMath-Meta datasets, with final accuracies of 65.25% and 42.36% on the respective datasets.

Table 2: The accuracy of various models across different datasets. Additionally, an ablation study is presented, illustrating the sequential application of different techniques for solving geometry mathematics problems. IC, QA, and MT stand for adding the GeoMath-IC, GeoMath-QA, and GeoMath-Meta datasets for training respectively. We indicate whether providing few-shot samples in the inference phase using 'with ICL' and 'w/o ICL'.

Models	GeoQA+ (%,↑)	GeoMath (%,↑)
Openflamingo-9B [5]	27.37	21.14
LLaVA-1.5-13B [22]	29.30	22.28
Bard [17]	47.10	20.00
GPT-4V [1]	50.50	20.00
DPE-NGS [9]	66.09	-
G-llava-13B [16]	**67.00**	-
Geo-LLaVA-13B (QA)	57.70	28.60
Geo-LLaVA-13B (IC+QA)	61.04 (+3.34)	37.12 (+8.52)
Geo-LLaVA-13B (IC+QA+MT, w/o ICL)	63.13 (+2.09)	41.48 (+4.36)
Geo-LLaVA-13B (IC+QA+MT, with ICL)	65.25 (+2.12)	**42.36** (+0.88)

The results indicate that the proposed Geo-LLaVA-13B model outperforms several other models in solving geometry mathematics problems, particularly when using a combination of the GeoMath-IC, GeoMath-QA, and GeoMath-Meta datasets on the solid geometric problems. The step-by-step improvements highlight the effectiveness of sequentially incorporating these datasets and techniques. Additionally, the model's performance further benefits from ICL, suggesting that providing few-shot examples during inference significantly enhances its accuracy.

5 Conclusion

In conclusion, this paper presents a novel approach to address the challenges inherent in multi-modal math problem-solving on geometry. Our research emphasizes the pivotal role of integrating meta-learning into models, enabling them to accurately interpret and reason through complex visual and textual inputs—an essential capability for resolving geometric problems effectively. Recognizing the limitations posed by the small size and the absence of reasoning steps in current geometry datasets, we have developed a pioneering dataset called GeoMath. This dataset amalgamates reasoning steps drawn from pre-existing datasets and materials sourced from a Chinese high school educational website, thereby filling a critical gap in available resources.

This study not only provides significant contributions to the domain of multi-modal geometry problem-solving but also paves the way for future research endeavors. The application of LLMs and LMMs to this domain promises to unlock new potential and methodologies. The development of GeoMath stands as a notable milestone, offering robust strategies for addressing the complexities of geometry problems. Future research could explore refining these models further, expanding the dataset with additional problem types, and investigating their applicability across various educational contexts, potentially transforming how multi-modal mathematical reasoning is approached in both academic and practical settings.

References

[1] Gpt-4v(ision) system card.

[2] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[3] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems 35* (2022), 23716–23736.

[4] Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403* (2023).

[5] Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390* (2023).

[6] Azerbayev, Z., Schoelkopf, H., Paster, K., Santos, M. D., McAleer, S., Jiang, A. Q., Deng, J., Biderman, S., and Welleck, S. Llemma: An open language model for mathematics. *arXiv preprint arXiv:2310.10631* (2023).

[7] Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966* (2023).

[8] Bordes, F., Pang, R. Y., Ajay, A., Li, A. C., Bardes, A., Petryk, S., Mañas, O., Lin, Z., Mahmoud, A., Jayaraman, B., Ibrahim, M., Hall, M., Xiong, Y., Lebensold, J., Ross, C., Jayakumar, S., Guo, C., Bouchacourt, D., Al-Tahan, H., Padthe, K., Sharma, V., Xu, H., Tan, X. E., Richards, M., Lavoie, S., Astolfi, P., Hemmat, R. A., Chen, J., Tirumala, K., Assouel, R., Moayeri, M., Talattof, A., Chaudhuri, K., Liu, Z., Chen, X., Garrido, Q., Ullrich, K., Agrawal, A., Saenko, K., Celikyilmaz, A., and Chandra, V. An introduction to vision-language modeling. Issue: arXiv:2405.17247.

[9] Cao, J., and Xiao, J. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th International Conference on Computational Linguistics* (2022), pp. 1511–1520.

[10] Chen, J., Li, T., Qin, J., Lu, P., Lin, L., Chen, C., and Liang, X. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. *arXiv preprint arXiv:2212.02746* (2022).

[11] Chen, J., Tang, J., Qin, J., Liang, X., Liu, L., Xing, E. P., and Lin, L. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *arXiv preprint arXiv:2105.14517* (2021).

[12] Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[13] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[14] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[15] Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning* (2017), PMLR, pp. 1126–1135.

[16] Gao, J., Pi, R., Zhang, J., Ye, J., Zhong, W., Wang, Y., Hong, L., Han, J., Xu, H., Li, Z., et al. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370* (2023).

[17] Google. Google bard. https://bard.google.com, 2024. Accessed: 2024-07-05.

[18] Ha, D., Dai, A. M., and Le, Q. V. Hypernetworks. In *ICLR* (2022).

[19] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[20] Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning* (2023), PMLR, pp. 19730–19742.

[21] Liang, Z., Yang, T., Zhang, J., and Zhang, X. Unimath: A foundational and multimodal mathematical reasoner. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (2023), pp. 7126–7133.

[22] Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems 36* (2024).

[23] Masry, A., Kavehzadeh, P., Do, X. L., Hoque, E., and Joty, S. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. *arXiv preprint arXiv:2305.14761* (2023).

[24] Min, S., Lewis, M., Zettlemoyer, L., and Hajishirzi, H. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943* (2021).

[25] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning* (2021), PMLR, pp. 8748–8763.

[26] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *ICML* (2021), vol. 139 of *Proceedings of Machine Learning Research*, PMLR, pp. 8748–8763.

[27] Ravi, S., and Larochelle, H. Optimization as a model for few-shot learning. In *ICLR* (2016).

[28] Schmidhuber, J. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook.* PhD thesis, Technische Universität München, 1987.

[29] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[30] Trinh, T. H., Wu, Y., Le, Q. V., He, H., and Luong, T. Solving olympiad geometry without human demonstrations. *Nature 625*, 7995 (2024), 476–482.

[31] Wang, K., Ren, H., Zhou, A., Lu, Z., Luo, S., Shi, W., Zhang, R., Song, L., Zhan, M., and Li, H. Mathcoder: Seamless code integration in llms for enhanced mathematical reasoning. *arXiv preprint arXiv:2310.03731* (2023).

[32] Weng, Y., Zhu, M., Xia, F., Li, B., He, S., Liu, S., Sun, B., Liu, K., and Zhao, J. Large language models are better reasoners with self-verification. *arXiv preprint arXiv:2212.09561* (2022).

[33] Zhang, M.-L., Yin, F., and Liu, C.-L. A multi-modal neural geometric solver with textual clauses parsed from diagram. *arXiv preprint arXiv:2302.11097* (2023).

[34] Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592* (2023).

SynthDoc: Bilingual Documents Synthesis for Visual Document Understanding

Chuanghao Ding[*] [†]
State Key Laboratory for Novel
Software Technology
Nanjing University
SenseTime Research
Nanjing, China
ch777.ding@smail.nju.edu.cn

Xuejing Liu[*]
SenseTime Research
Shanghai, China
xuejing931210@gmail.com

Wei Tang[†]
SenseTime Research
Shanghai, China
weitang@njust.edu.cn

Juan Li
State Key Laboratory for Novel
Software Technology
Nanjing University
Nanjing, China
juanli@smail.nju.edu.cn

Xiaoliang Wang
State Key Laboratory for Novel
Software Technology
Nanjing University
Nanjing, China
waxili@nju.edu.cn

Rui Zhao
SenseTime Research
Shanghai, China
zhaorui@sensetime.com

Cam-Tu Nguyen[‡]
State Key Laboratory for Novel
Software Technology
School of Artificial Intelligence
Nanjing University
Nanjing, China
ncamtu@nju.edu.cn

Fei Tan[‡]
SenseTime Research
Shanghai, China
tanfei2007@gmail.com

Abstract

This paper introduces SynthDoc, a novel synthetic document generation pipeline designed to enhance Visual Document Understanding (VDU) by generating high-quality, diverse datasets that include text, images, tables, and charts. Addressing the challenges of data acquisition and the limitations of existing datasets, SynthDoc leverages publicly available corpora and advanced rendering tools to create a comprehensive and versatile dataset. Our experiments, conducted using the Donut model, demonstrate that models trained with SynthDoc's data achieve superior performance in pre-training read tasks and maintain robustness in downstream tasks, despite language inconsistencies. The release of a benchmark dataset comprising 5,000 image-text pairs not only showcases the pipeline's capabilities but also provides a valuable resource for the VDU community to advance research and development in document image recognition. This work significantly contributes to the field by offering a scalable solution to data scarcity and by validating the efficacy of end-to-end models in parsing complex, real-world documents.

CCS Concepts

• **Applied computing → Multi / mixed media creation**.

Keywords

Visual Document Understanding, End-to-End Document Parsing, Synthetic Document Generation

ACM Reference Format:
Chuanghao Ding, Xuejing Liu, Wei Tang, Juan Li, Xiaoliang Wang, Rui Zhao, Cam-Tu Nguyen, and Fei Tan. 2024. SynthDoc: Bilingual Documents Synthesis for Visual Document Understanding. 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, 10 pages. https://doi.org/10.1145/3688866.3689125

1 Introduction

Visual Document Understanding (VDU) is a complex endeavor that seeks to decipher and interpret information from documents across a spectrum of formats and layouts [11, 16, 25, 37, 42]. The objective of VDU is to develop algorithms capable of grasping the content, structure, and context of documents, thereby enabling tasks such as document classification [11], text detection [16, 35], layout analysis [37, 42], and object detection [24, 25].

[*] Equal contribution to this research.
[†] Work was done during internship at SenseTime Research.
[‡] Corresponding author

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

ACM ISBN 979-8-4007-1193-0/24/10
https://doi.org/10.1145/3688866.3689125

Current research in VDU predominantly employs two methodologies: one [2, 13, 26, 30, 47, 48] relies on OCR technology to convert document images into text for subsequent processing, while the other [1, 5, 17, 20, 28, 32, 33] adopts an end-to-end approach, analyzing the document images directly. The pre-training and fine-tuning paradigm is extensively utilized in multimodal learning [5, 9, 17, 20, 29, 41]. The end-to-end approach leverages this paradigm to incorporate robust text recognition capabilities into the model, addressing the limitations of OCR accuracy and achieving high processing efficiency. A common pre-training task is the text reading task, and previous studies [17, 20] have demonstrated its efficacy in enhancing model performance across various downstream tasks, such as document parsing and document Visual Question Answering (VQA). Therefore, leveraging the text reading task to bolster the capabilities of the base model is of paramount importance.

The data requirements for the text reading task encompass two main aspects: high-quality document images and corresponding text annotations that reflect the reading order. Obtaining such data is intricate, with existing methods either depending on large-scale public document datasets and additional OCR models to generate pseudo-labels [17] or relying on complex data processing pipelines to scrape document data from the web [43]. However, these methods often result in low-quality labels, face copyright restrictions, and contend with data noise. Moreover, they typically focus only on specific elements within document images, such as text or certain document components. For example, Nougat [5] and KOSMOS-2.5 [32] concentrate on table parsing, while MatCha [28] emphasizes chart rendering. It is rare to find a dataset that encompasses all document elements simultaneously. A recent approach Vary [44], while employing rendering of various document types, has utilized only over 10 templates, which falls short in terms of the richness of document layouts.

To tackle the limitations in document layout richness and the challenges associated with data acquisition, we introduce SynthDoc, a synthetic document generation pipeline. This pipeline is designed to create datasets that include text, images, tables, and a variety of charts. We begin by aggregating publicly available datasets, which have been validated on large language or multimodal models, to form our text and image corpora. We then enhance the TableGeneration [46] to produce a diverse set of tables, and use tools like pandas [34], Matplotlib [14], and ECharts [23] to generate chart-table pairs, thus expanding our chart data corpus. Therefore, our approach provides three distinct advantages: 1) Synthdoc can leverage redundant, open-resources NLP datasets to generate high-resolution, coherent content for multimodal model training. 2) Synthdoc is developed with high efficiency, precision, and dynamically customizes document layouts and features robust scalability. 3) The synthesized data include comprehensive content and structural annotations, facilitating the pre-training of structured document parsing models based on LLMs. Synthetic data can effectively complement the expensive manually labeled real datasets.

Our comprehensive experiments, leveraging the Donut model, have yielded compelling results that underscore the efficacy of the SynthDoc pipeline. The models trained with our synthesized document images have achieved remarkable accuracy in the pre-training read task, demonstrating a keen ability to parse both Chinese and English text, as well as tables and charts within the generated datasets. This proficiency extends to the fine-tuning phase of downstream tasks, where the models maintain a high level of performance despite the primary and secondary tasks involving languages that are not always consistent.

Furthermore, we have conducted visual analyses of the models' parsing capabilities on more complex, real-world documents. Despite the relatively limited variety of document types synthesized by our pipeline, the models have shown commendable results in parsing these intricate documents. A particularly surprising finding pertains to the chart parsing capabilities. In instances where scatter plots did not explicitly label the x-axis, our models were able to accurately infer the horizontal coordinates. This suggests that the models trained with our rendered data possess a certain level of spatial understanding and an awareness of the sequence among numerical values.

In response to the absence of comprehensive public datasets for model validation in document image parsing, we have released a set of 5,000 images based on the SynthDoc pipeline. This release not only showcases the quality and diversity of the document data we generate but also provides a benchmark for the document image recognition community to advance and develop new methodologies.

In summary, the key contributions of this paper are as follows:

- SynthDoc Pipeline: We introduce a novel synthetic data pipeline for document images, named SynthDoc, which utilizes publicly available text or text-image pairs along with rendered tables and charts. This pipeline is capable of simultaneously generating text, images, tables, and various types of charts within document images.
- Benchmark Release: We have made available to the research community a benchmark dataset consisting of 5,000 image-text pairs. This release aims to highlight the robustness of the data produced by our pipeline and to support further research and development in the area of document image parsing.
- Experimental Validation: Through experiments based on the Donut model, we have demonstrated that our proposed dataset and training methodology lead to a significant enhancement in the model's document image parsing capabilities. Additionally, the models trained with our approach maintain competitive performance across a range of downstream tasks.

2 Related Work

2.1 Image Document Data

Deep learning-based document image understanding has consistently been recognized as a significant and intricate work, and many datasets have been proposed to parse and understand document images from different perspectives. For example, FUNSD [16] is utilized for form understanding. RVL-CDIP [11]is employed for document classification. PubLayNet [52]is utilized for document layout analysis in our study. However, these datasets fail to meet the requirements of recent end-to-end methods, which rely on

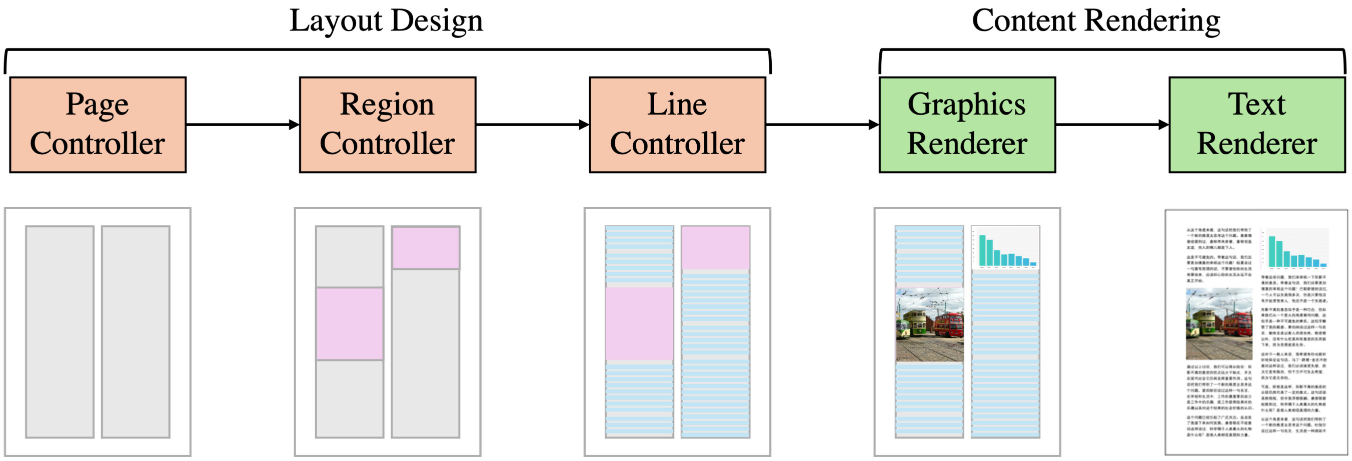


Figure 1: The pipeline of Document Image Synthesis, including layout design and content rendering. The layout design involves planning at three scales: full-page, regional, and line-by-line. Content rendering creates both visual graphics and textual content.

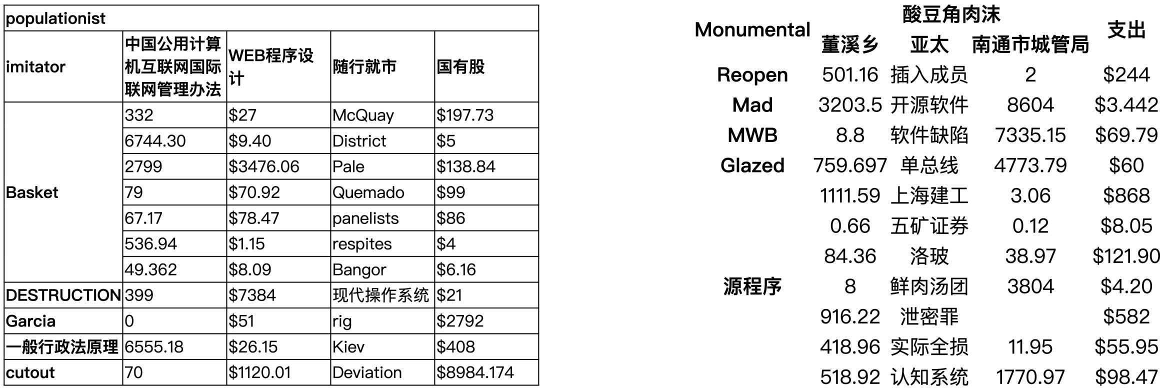
large amounts of document image data for pretraining. Some approaches [6, 17] parse existing document datasets, such as IIT-CDIP [21], by commercial OCR models. However the quality of datasets obtained by such methods is constrained by OCR accuracy, and utilizing commercial OCR models can be costly. Another approachs [32, 42, 43] rely on the crawler techniques to collect extensive data from the internet, extracting document image data through parsing and filtering, which often yield datasets with considerable noise, due to the complexity of the document, and are subject to copyright restrictions. Unlike these methods, we collect existing web-scale datasets [8, 18, 36, 39, 40, 45, 49] that have been used by large language models or multimodal large language models, employing a synthetic approach to obtain document image data, which can yield clean data and include complex elements such as charts.

2.2 Text Reading Task

As the end-to-end multimodal model evolves, the task of text reading within document images has gained increasing attention from scholars, affirming its significant value in the field. For example, Donut [17] is pre-trained on document images and their associated text annotations, reading text from images one by one according to previous text contexts. Nougat [5] follows Donut's model and training approach, with a specialized focus on the domain of scientific papers, adeptly reading texts, tables, and formulas using markup language. DocParser [33] introduces the Masked Document Reading method, which is designed to enhance the model's reasoning capabilities by predicting the text situated within the masked regions. UReader [50] utilizes text reading task to train multimodal large language model, and proposes to predict text from any position of document images, which ensures the model can read different parts of texts with the context. Pix2struct [20] found that the text reading task showed a strong curriculum learning effect, using it as warmup phase resulted in more stable training, faster convergence, and better performance. It is worth noting that all of these tasks require millions of document images, kosmos2.5 [32] collected 324.4M data from public datasets and web, such as IIT-CDIP [21], arxiv, and GitHub. However, these data are difficult to obtain and have copyright restrictions, so we propose a data rendering pipeline for text reading task to improve the model's understanding of dual-language documents.

2.3 Synthetic Document Image

Document image data generation has been widely concerned in the field of visual document understanding. Some document image generation algorithms, based on GAN networks, generate plausible document images, emphasizing the diversity and quality of generated documents. For example, Biswas et al.[3] utilize the GAN model to generate diverse and credible document images based on the provided layout. Zheng et al.[51] proposed a layout depth generation model for graphic design, which implicitly captured the influence of visual and text content on layout, and synthesized complex layout design according to the visual and text semantics input by users. However, these methods do not consider the annotation information used for visual document understanding, the quality and size of the generated images are limited by the model, and additional training models are required for different languages, which is inefficient. Other methods generate document and ground truth pairs for specific visual document understanding tasks. For document layout analysis, Pisaneschi et al.[38] generates document layout information based on LayoutTransformer[10] and additional post-processing methods which fill in the corresponding texts, images, and Mathematical objects based on the model or the collected corpus. Ling et al.[27] proposed the document domain randomization approach to simulate the document layout, and then randomly fill in collected elements such as texts and images. For pretraining of Document Intelligence tasks, Biten et al.[4] generates large-scale pre-training datas with OCR annotation information on IDL datasets based on commercial OCR tools. However, the current pre-training of intelligent document understanding based on



populationist				
imitator	中国公用计算机互联网国际联网管理办法	WEB程序设计	随行就市	国有股
Basket	332	$27	McQuay	$197.73
	6744.30	$9.40	District	$5
	2799	$3476.06	Pale	$138.84
	79	$70.92	Quemado	$99
	67.17	$78.47	panelists	$86
	536.94	$1.15	respites	$4
	49.362	$8.09	Bangor	$6.16
DESTRUCTION	399	$7384	现代操作系统	$21
Garcia	0	$51	rig	$2792
一般行政法原理	6555.18	$26.15	Kiev	$408
cutout	70	$1120.01	Deviation	$8984.174

(a)

Monumental	酸豆角肉沫			支出
	董溪乡	亚太	南通市城管局	
Reopen	501.16	插入成员	2	$244
Mad	3203.5	开源软件	8604	$3.442
MWB	8.8	软件缺陷	7335.15	$69.79
Glazed	759.697	单总线	4773.79	$60
	1111.59	上海建工	3.06	$868
	0.66	五矿证券	0.12	$8.05
	84.36	洛玻	38.97	$121.90
源程序	8	鲜肉汤团	3804	$4.20
	916.22	泄密罪		$582
	418.96	实际全损	11.95	$55.95
	518.92	认知系统	1770.97	$98.47

(b)

Figure 2: Gridlined and gridless table renderings.

large language models relies on document image parsing tasks, and the existing data can no longer meet the training demands, so we propose a new data set generation pipeline to synthesize accurate, clear, logical and coherent document parsing datasets to adapt to the development of visual document understanding.

A similar effort to this paper is donut[17], which uses a portion of generated data to supplement data in different languages. The difference is that their work randomly pastes text into images and ignores layout information and structured elements such as tables, charts and images.

3 Document Image Synthesis

In this section, we delve into the pipeline for generating document images, which is primarily composed of two key components: layout design and content rendering, as shown in Fig. 1. The layout design encompasses the architectural planning at three distinct scales: the entire page, individual regions, and lines of text. This meticulous arrangement ensures that the document's structure aligns with conventional reading habits while maximizing visual diversity. Content rendering, on the other hand, is responsible for the creation of both graphic and textual elements. This phase includes the rendering of graphics, which can consist of tables, images, and charts, as well as the rendering of text. Each element is crafted with attention to detail, ensuring that the final document image not only conveys information accurately but also presents it in an aesthetically pleasing and reader-friendly manner.

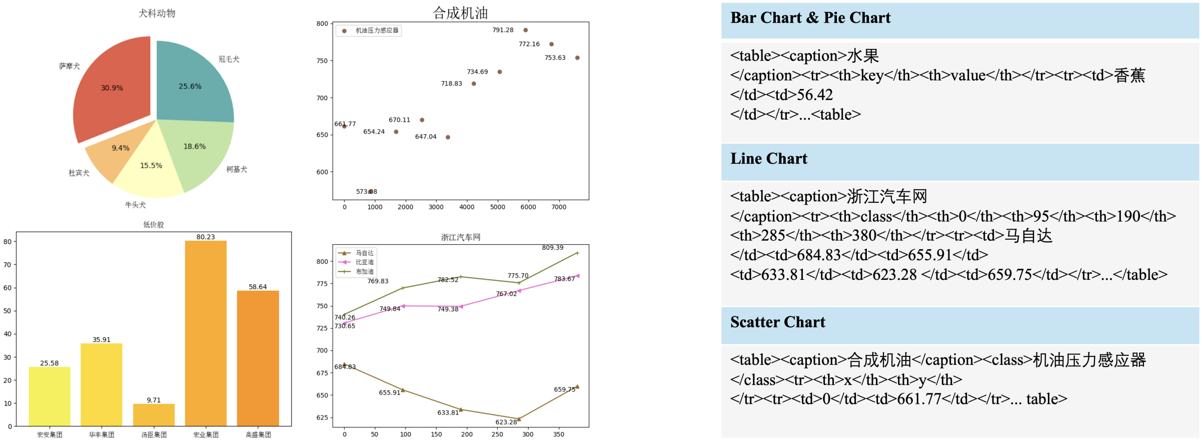
3.1 Layout Design

The document image synthesis pipeline comprises three integral components: the Page Controller, Region Controller, and Line Controller. The Page Controller ensures a consistent and visually appealing layout by defining and maintaining layout elements and typographical attributes. The Region Controller segments the document into distinct areas for various content types, facilitating a logical and balanced composition. Lastly, the Line Controller meticulously organizes text, applying typographical rules to enhance readability and engagement. Together, these components work to create structured, professional-looking documents that are both informative and aesthetically pleasing.

3.1.1 Page Controller. The Page Controller is instrumental in establishing a consistent and visually appealing layout for single-page documents. It sets and maintains the uniformity of layout elements such as data areas, page margins, and the spacing between segments and lines. Additionally, it oversees the font size and color palette, ensuring that the document's visual presentation is coherent and reader-friendly. This component's role is critical in creating a structured and professional look that enhances the document's overall readability and impact.

3.1.2 Region Controller. The Region Controller plays a pivotal role in the document's structural integrity by meticulously segmenting the data areas into distinct regions for text, images, tables, and charts. It operates on a macro level, determining where each type of content will be placed to optimize readability and visual impact. This controller ensures that the document's layout supports a logical flow, with areas designated for complex data representations such as charts and tables, and separate sections for textual content. By carefully allocating space for each element, the Region Controller ensures that the document's overall composition is balanced and adheres to the principles of good document design, allowing readers to navigate the information with ease.

3.1.3 Line Controller. The Line Controller is responsible for the micro-level organization of textual content within the document. It takes the individual word images produced by the Text Renderer and arranges them into coherent lines, respecting the predefined attributes such as word spacing, line height, and alignment. This controller's work is crucial for establishing the document's typographical style, which includes setting the rhythm and pacing of the text. By fine-tuning the line breaks, indentations, and other typographical elements, the Line Controller ensures that the text is not only legible but also visually engaging. This attention to detail in formatting contributes to a professional and polished appearance, enhancing the document's overall presentation quality.



Bar Chart & Pie Chart

```
<table><caption>水果
</caption><tr><th>key</th><th>value</th></tr><tr><td>香蕉
</td><td>56.42
</td></tr>...<table>
```

Line Chart

```
<table><caption>浙江汽车网
</caption><tr><th>class</th><th>0</th><th>95</th><th>190</th>
<th>285</th><th>380</th></tr><tr><td>马自达
</td><td>684.83</td><td>655.91</td>
<td>633.81</td><td>623.28 </td><td>659.75</td></tr>...</table>
```

Scatter Chart

```
<table><caption>合成机油</caption><class>机油压力感应器
</class><tr><th>x</th><th>y</th>
</tr><tr><td>0</td><td>661.77</td></tr>... table>
```

(a)

(b)

Figure 3: (a) Samples of the synthesized charts: Pie Chart, Vertical Bar Chart, Scatter Chart and Line Chart. (b) The annotation formats corresponding to different charts, which are presented in HTML format.

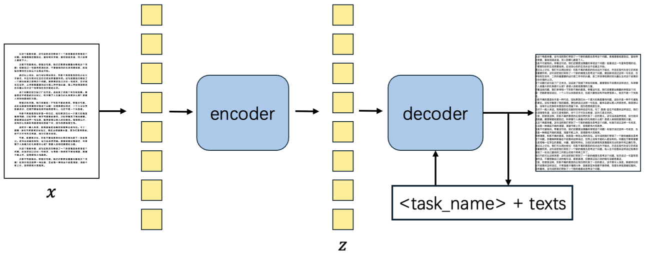


Figure 4: This is an overview architecture to training the model

3.2 Content Rendering

With the layout meticulously established, the pipeline transitions to the content rendering phase, where the visual and textual elements of the document come to life. This stage involves the intricate process of integrating graphics and text, ensuring that each component not only complements the layout but also enhances the document’s overall narrative and aesthetic appeal.

3.3 Experimental Results

3.3.1 Graphic Renderer. The Graphic Renderer is a sophisticated component of our pipeline, dedicated to the rendering of images, tables, and charts. For images, we focus on incorporating natural images, where available category data is used to caption and embed the images within the document. If category information is present, the returned text represents the category; otherwise, it is replaced with a generic placeholder "<nature_image>". This approach ensures that each image is contextually relevant and enhances the document’s informational content.

In the realm of tables, we have designed two distinct types to accommodate various data presentations. The first type features complete borders, suitable for complex data with line breaks within cells, while the second type adopts a minimalist or borderless style, aligning with the prevalent aesthetic in research publications. Both types incorporate random cell merging to manage data complexity effectively. The rendered tables are displayed in Fig. 2.

For charts, our pipeline supports the rendering of four chart types: bar, pie, line, and scatter plots. Bar charts, available in both horizontal and vertical orientations, are crafted for data comparison, with key-value pairs represented in a tabular format to facilitate readability. To mitigate issues with overlapping labels in vertical bar charts, we implement random fonts and rotation angles. Pie charts, similar in rendering to bar charts, require that the aggregated values represent a total of 1 or 100, expressed as decimals or percentages. Line charts illustrate trends over time or variables, with each chart featuring a unique set of data groups and points, generating an image-label pair. Scatter plots, used to depict the distribution of a single element, employ a label and x and y coordinates for each point, with the number of points limited to a range of [5, 20] to manage complexity. The generated examples are depicted in Fig. 3a. The corresponding HTML annotations are displayed in Fig. 3b.

The emphasis on the model’s ability to understand the structure of diverse elements is paramount. We refrain from using AI tools to generate data within elements, instead leveraging an open textual corpus for our tables and charts, ensuring the authenticity and relevance of the data. The matplotlib library is utilized for chart rendering, and we have refined table rendering techniques to better integrate with the document’s overall design.

3.3.2 Text Renderer. The Text Renderer plays an indispensable role in the content rendering process, meticulously generating word images for each word in the text. This method affords a high level of control over the typography and layout, ensuring that the text is not only legible but also aesthetically integrated with the document’s visual elements. The Text Renderer works in concert with the Graphic Renderer to weave a cohesive and engaging narrative,

Table 1: The comparison between different methods across diverse synthetic documents.

Metrics	Methods	Pure Document		Complex Document			Average
		English	Chinese	Doc w/image	Doc w/table	Doc w/chart	
AED ↓	Donut [17]	0.3764	0.5148	0.7631	0.8679	0.9097	0.6864
	vary [44]	0.1452	0.1760	0.5598	0.7415	0.6663	0.4578
	our	**0.0321**	**0.1370**	**0.1665**	**0.0583**	**0.1029**	**0.0994**
F1-score ↑	Donut [17]	0.9370	0.8107	0.3720	0.4573	0.2840	0.5722
	vary [44]	0.8554	0.9002	0.5852	0.5854	0.6531	0.7159
	our	**0.9611**	**0.9020**	**0.8855**	**0.9199**	**0.8810**	**0.9099**
Prediction ↑	Donut [17]	0.9534	0.8256	0.4061	0.5302	0.4063	0.6243
	vary [44]	0.8762	0.8974	0.6383	0.7026	0.7961	0.7821
	our	**0.9717**	**0.9136**	**0.9065**	**0.9347**	**0.9017**	**0.9256**
Recall ↑	Donut [17]	0.9228	0.8015	0.3647	0.4313	0.2540	0.5549
	vary [44]	0.8482	**0.9044**	0.5746	0.5501	0.5868	0.6928
	our	**0.9515**	0.8916	**0.8682**	**0.9076**	**0.8636**	**0.8965**

Table 2: Performance Comparison of different methods on CORD.

Model	OCR	Acc	Precision	Recall	F1
BERT [15]	√	78.2	-	-	82.2
BROS [12]	√	80.3	-	-	83.7
LayoutLMv2 [48]	√	87.0	-	-	88.9
KOSMOS-2.5 [32]	-	-	83.64	87.83	85.69
Donut [17]	-	93.5	-	-	91.6
our	-	90.1	82.6	83.3	82.9

blending visual and textual information to enhance the reader's experience.

Following Donut's data generation approach, the Text Renderer creates a word image for each word, which is crucial for the document's visual composition and label generation. This attention to detail in text rendering ensures that the document's textual content is as carefully crafted as its visual elements, contributing to a polished and professional final product.

3.4 Concerns of Data Generation Pipeline

3.4.1 Scalability. Even if we generate as much diverse data as possible, it hardly covers all real-world document layouts. To mitigate this, we've integrated real document images into our benchmark to maximize layout variability. However, it is worth noting that our solution is highly adaptable, with scalability in two key dimensions: 1) **Layout Customization:** We allow for tailored document layouts to swiftly and cost-effectively expand our training data to fit various scenarios. 2) **Language Independence:** Our pipeline transcends language barriers, enabling document image generation in any language. For instance, we've produced French documents using the ROOTS[19] dataset.

3.4.2 Data Privacy. Our pipeline allows for local regulatory adaptation and reproducibility of datasets through customizable pipeline components. We advocate for the use of public corpora and tools to foster transparency and verifiability in research.

4 Training on SynthDoc

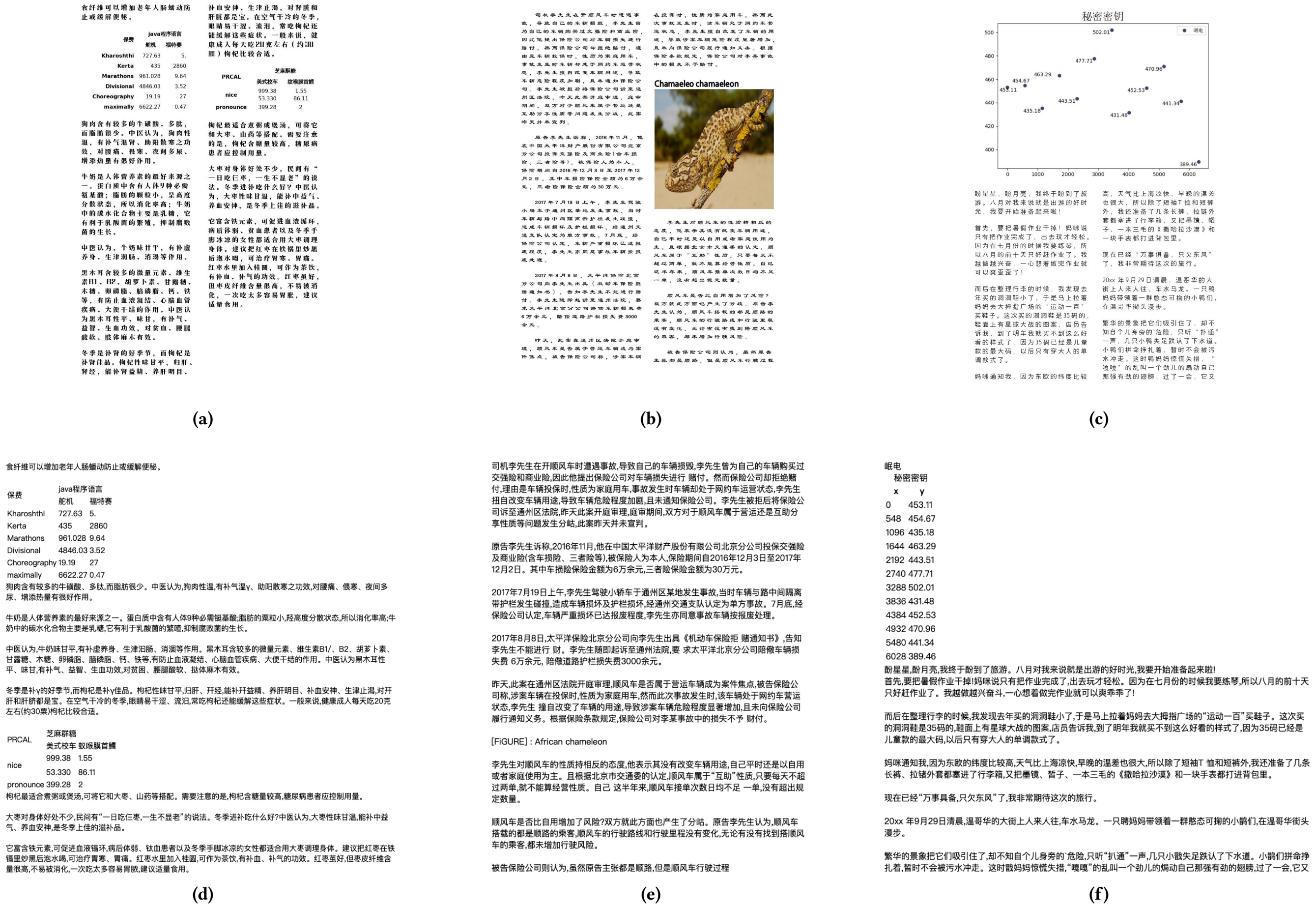
This section details the pre-training of the model based on the Donut architecture, focusing on its parsing performance with bilingual (English and Chinese) documents. The primary objective is to validate the model's ability to effectively handle and interpret content in both languages, ensuring its suitability for multilingual document analysis.

4.1 Model Architecture

Unlike previous OCR-based approaches [2, 13] for visual document understanding tasks, recent research [20, 33] has shifted towards parsing document images in an end-to-end fashion, eliminating the need for OCR results as input. The dataset we generated primarily aims to enhance and validate the visual document parsing capabilities of this end-to-end models. Illustrated in Figure 4, our model is constructed based on the Donut architecture. We follow the Donut [17], utilizing the Swin-Transformer [31] as our visual encoder. Previous experiments have demonstrated its superior performance compared to ViT [7]. We employ mBART [22] as the decoder, which has stronger noise robustness and multilingual capabilities.

4.2 Implementation Details

Following the previous works [5, 17], we employ Swin-Base as the encoder and the first four layers of mBART as the decoder, with a patch size of 4 and a window size of 10. We set the input image size to (H, W) = (1280, 960) to meet the requirements of Swin-Base for image dimensions. For pre-training, we set a batch size of 192 and employ the AdamW optimizer, initializing the learning rate at 5e-5 and setting a minimum of 7.6e-6, while utilizing an exponential scheduler with a gamma of 0.9996, updating the learning rate every 16 training steps. For fine-tuning, we utilize a cosine scheduler with



(a) (b) (c)

食纤维可以增加老年人肠蠕动防止或缓解便秘。

保费	java程序语言	
	舵机	福特赛
Kharoshthi	727.63	5.
Kerta	435	2860
Marathons	961.028	9.64
Divisional	4846.03	3.52
Choreography	19.19	27
maximally	6622.27	0.47

狗肉含有较多的牛磺酸、多肽,而脂肪很少。中医认为,狗肉性温,有补气温γ、助阳散寒之功效,对腰痛、偎寒、夜间多尿、增添热量有很好作用。

牛奶是人体营养素的最好来源之一。蛋白质中含有人体9种必需铤基酸;脂肪的粟粒小,羟高度分散状态,所以消化率高;牛奶中的碳水化合物主要是乳糖,它有利于乳酸菌的繁喳,抑制腐败菌的生长。

中医认为,牛奶味甘平,有补虚养身、生津汩肠、消涸等作用。黑木耳含较多的微量元素、维生素B1/、B2、胡芗卜素、甘露糖、木糖、卵磷脂、脑磷脂、钙、铁等,有防止血液凝结、心脑血管疾病、大便干结的作用。中医认为黑木耳性平、味甘,有补气、益智、生血功效,对贫困、腰腿酸软、挞体麻木有效。

冬季是补γ的好季节,而枸杞是补γ佳品。枸杞性味甘平,归肝、幵经,能补幵益精、养肝明目、补血安神、生津止渴,对幵肝和肝脐都是宝。在空气干冷的冬季,眼睛易干涩、流汩,常吃枸杞还能缓解这些症状。一般来说,健康成人每天吃20克左右(约30粟)枸杞比较合适。

PRCAL	芝麻群糖	
	美式校车	蚁喉膜首鳕
nice	999.38	1.55
	53.330	86.11
pronounce	399.28	2

枸杞最适合煮粥或煲汤,可将它和大枣、山药等搭配。需要注意的是,枸杞含糖量较高,糖尿病患者应控制用量。

大枣对身体好处不少,民间有"一日吃仨枣,一生不显老"的说法。冬季进补吃什么好?中医认为,大枣性味甘温,能补中益气、养血安神,是冬季上佳的滋补品。

它富含铁元素,可促进血液镐环,病后体弱、钛血患者以及冬季手脚冰凉的女性都适合用大枣调理身体。建议把红枣在铁镉里炒黑后泡水喝,可治疗胃寒、胃痛。红枣水里加入桂圆,可作为茶饮,有补血、补气的功效。红枣虽好,但枣皮纤维含量很高,不易被消化,一次吃太多容易胃脓,建议适量食用。

(d)

司机李先生在开顺风车时遭遇事故,导致自己的车辆损毁,李先生曾为自己的车辆购买过交强险和商业险,因此他提出保险公司对车辆损失进行 赌付。然而保险公司却拒绝赌付,理由是车辆投保时,性质为家庭用车,事故发生时车辆却处于网约车运营状态,李先生扭自改变车辆用途,导致车辆危险程度加剧,且未通知保险公司。李先生被拒后将保险公司诉至通州区法院,昨天此案开庭审理,庭审期间,双方对于顺风车属于营运还是互助分享性质等问题发生分岵,此案昨天并未宣判。

原告李先生诉称,2016年11月,他在中国太平洋财产股份有限公司北京分公司投保交强险及商业险(含车损险、三者险等),被保险人为本人,保险期间自2016年12月3日至2017年12月2日。其中车损险保险金额为6万余元,三者险保险金额为30万元。

2017年7月19日上午,李先生驾驶小轿车于通州区某地发生事故,当时车辆与路中间隔离带护栏发生碰撞,造成车辆损坏及护栏损坏,经通州交通支队认定为单方事故。7月底,经保险公司认定,车辆严重损坏已达报废程度,李先生亦同意事故车辆按报废处理。

2017年8月8日,太平洋保险北京分公司向李先生出具《机动车保险拒 赌通知书》,告知李先生不能进行 财。李先生随即起诉至通州法院,要 求太平洋北京分公司陪偸车辆损失费 6万余元, 陪偸道路护栏损失费3000余元。

昨天,此案在通州区法院开庭审理,顺风车是否属于营运车辆成为案件焦点,被告保险公司称,涉案车辆在投保时,性质为家庭用车,然而此次事故发生时,该车辆处于网约车营运状态,李先生 擅自改变了车辆的用途,导致涉案车辆危险程度显著增加,且未向保险公司履行通知义务。根据保险条款规定,保险公司对李某事故中的损失不予 财付。

[FiGURE] : African chameleon

李先生对顺风车的性质持相反的态度,他表示其没有改变车辆用途,自己平时还是以自用或者家庭使用为主。且根据北京市交通委的认定,顺风车属于"互助"性质,只要每天不超过两单,就不能算经营性质。自己 这半年来,顺风车接单次数日均不足 一单,没有超出规定数量。

顺风车是否比自用增加了风险?双方就此方面也产生了分岵。原告李先生认为,顺风车搭载的都是顺路的乘客,顺风车的行驶路线和行驶里程没有变化,无论有没有找到搭顺风车的乘客,都未增加行驶风险。

被告保险公司则认为,虽然原告主张都是顺路,但是顺风车行驶过程

(e)

岷电

秘密密钥

x	y
0	453.11
548	454.67
1096	435.18
1644	463.29
2192	443.51
2740	477.71
3288	502.01
3836	431.48
4384	452.53
4932	470.96
5480	441.34
6028	389.46

酚星星,酚月亮,我终于酚到了旅游。八月对我来说就是出游的好时光,我要开始准备起来啦!
首先,要把暑假作业干掉!妈咪说只有把作业完成了,出去玩才轻松。因为在七月份的时候我要练琴,所以八月的前十天只好赶作业了。我越做越兴奋斗,一心想着做完作业就可以爽乖乖了!

而后在整理行李的时候,我发现去年买的洞洞鞋小了,于是马上拉着妈妈去大拇指广场的"运动一百"买鞋子。这次买的洞洞鞋是35码的,鞋面上有星球大战的图案,店员告诉我,到了明年我就买不到这么好看的样式了,因为35码已经是儿童款的最大码,以后只有穿大人的单调款式了。

妈咪通知我,因为东欧的纬度比较高,天气比上海凉快,早晚的温差也很大,所以除了短袖T 恤和短裤外,我还准备了几条长裤、拉锗外套都塞进了行李箱,又把墨镜、皙子、一本三毛的《撒哈拉沙漠》和一块手表都打进背包里。

现在已经"万事具备,只欠东风"了,我非常期待这次的旅行。

20xx 年9月29日清晨,温哥华的大街上人来人往,车水马龙。一只聘妈妈带领着一群憨态可掬的小鹊们,在温哥华街头漫步。

繁华的景象把它们吸引住了,却不知自个儿身旁的`危险,只听"扒通"一声,几只小戳失足跌认了下水道。小鹊们拼命挣扎着,暂时不会被污水冲走。这时戳妈妈惊慌失措,"嘎嘎"的乱叫一个劲儿的煽动自己那强有劲的翅膀,过了一会,它又

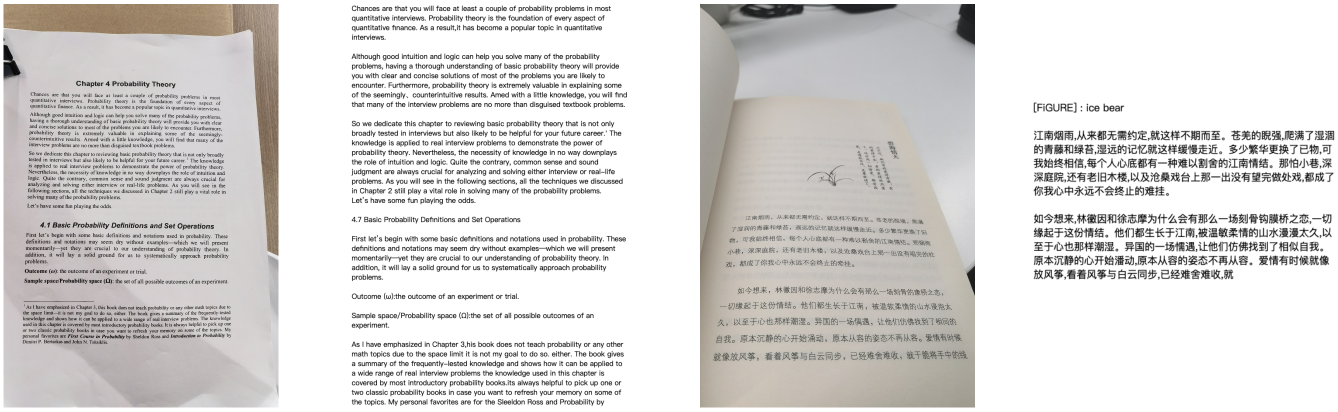
(f)

Figure 5: Examples of document image parsing on synthesized document with tables, images, and chart. (a), (b) and (c) stand for the synthetic document images with tables, images, and chart, (d), (e), and (f) represent the parsing results of the model on them, respectively.

a learning rate of 3e-5 to optimize our model, dynamically adjusting the input size according to the datasets, a practice effectively demonstrated by Donut.

4.2.1 Document Image Parsering. We evaluate the document parsing capabilities of other end-to-end models using the proposed benchmark in this paper and compare them with the performance of the model we trained. As shown in Table 1, we evaluate the models on five types of documents: English documents, Chinese documents, documents with natural images, documents with tables, and documents with charts. We observed that all models exhibited strong performance on English and Chinese documents, except for Donut, which showed slightly inferior results on the Average Edit Distance (AED), possibly due to its lack of training on the document dataset. However, with the exception of our model, all models displayed inadequate performance on complex documents containing additional elements. Specifically, our model achieved 0.1665, 0.0583, and 0.1029 AED on document images with images, tables, and charts, respectively, showing reductions of 0.3933, 0.6832, and 0.5634 compared to the Vary. It is noteworthy that in our benchmark, text labels associated with other elements represent only a small portion. This observation indicates that elements such as images in documents can significantly impact the model's text parsing capability.

4.2.2 Results on CORD.. The CORD dataset is a collection of data used for receipt recognition, comprising 800 samples for training and 100 samples for testing. Our pipeline's performance on the English CORD dataset did not demonstrate the expected improvements, due to the substantial distribution bias towards Chinese, which can be addressed by enhancing our model to more adeptly handle English-language documents in subsequent research. However, it is worth noting that our model not only improves its proficiency in Chinese document image recognition but also ensures comparable performance in downstream tasks.



Chances are that you will face at least a couple of probability problems in most quantitative interviews. Probability theory is the foundation of every aspect of quantitative finance. As a result,it has become a popular topic in quantitative interviews.

Although good intuition and logic can help you solve many of the probability problems, having a thorough understanding of basic probability theory will provide you with clear and concise solutions of most of the problems you are likely to encounter. Furthermore, probability theory is extremely valuable in explaining some of the seemingly、counterintuitive results. Amed with a little knowledge, you will find that many of the interview problems are no more than disguised textbook problems.

So we dedicate this chapter to reviewing basic probability theory that is not only broadly tested in interviews but also likely to be helpful for your future career.' The knowledge is applied to real interview problems to demonstrate the power of probability theory. Nevertheless, the necessity of knowledge in no way downplays the role of intuition and logic. Quite the contrary, common sense and sound judgment are always crucial for analyzing and solving either interview or real–life problems. As you will see in the following sections, all the techniques we discussed in Chapter 2 still play a vital role in solving many of the probability problems.
Let's have some fun playing the odds.

4.7 Basic Probability Definitions and Set Operations

First let's begin with some basic definitions and notations used in probability. These definitions and notations may seem dry without examples—which we will present momentarily—yet they are crucial to our understanding of probability theory. In addition, it will lay a solid ground for us to systematically approach probability problems.

Outcome (ω):the outcome of an experiment or trial.

Sample space/Probability space (Ω):the set of all possible outcomes of an experiment.

As I have emphasized in Chapter 3,his book does not teach probability or any other math topics due to the space limit it is not my goal to do so. either. The book gives a summary of the frequently–lested knowledge and shows how it can be applied to a wide range of real interview problems the knowledge used in this chapter is covered by most introductory probability books.its always helpful to pick up one or two classic probability books in case you want to refresh your memory on some of the topics. My personal favorites are for the Sleeldon Ross and Probability by

[FiGURE] : ice bear

江南烟雨,从来都无需约定,就这样不期而至。苍羌的睨强,爬满了湿涸的青藤和绿苔,湿远的记忆就这样缓慢走近。多少繁华更换了已物,可我始终相信,每个人心底都有一种难以割舍的江南情结。那怕小巷,深深庭院,还有老旧木楼,以及沧桑戏台上那一出没有望完做处戏,都成了你我心中永远不会终止的难挂。

如今想来,林徽因和徐志摩为什么会有那么一场刻骨钩膜桥之恋,一切缘起于这份情结。他们都生长于江南,被温敏柔情的山水漫漫太久,以至于心也那样潮湿。异国的一场懦遇,让他们仿佛找到了相似自我。原本沉静的心开始潘动,原本从容的姿态不再从容。爱情有时候就像放风筝,看着风筝与白云同步,已经难舍难收,就

(a) (b) (c) (d)

Figure 6: Examples of document image parsing on real English and Chinese documents. (a) real English document (b) prediction of English documents (c) real Chinese document (d) prediction of Chinese documents.

4.3 Visual Analysis

We provide sufficient visualization results of our model to demonstrate the excellent performance of the model in text image recognition. Specifically, The Figure 5 illustrates synthetic images, containing tables, images, and charts demonstrating our model's ability to parse text, tables, images, and charts information in a manner consistent with human reading order. Furthermore, as illustrated in the last row of Figure 6, our model exhibits robust parsing capability when applied to real document images.

4.3.1 Spatial Understanding. We observed that end-to-end models possess strong spatial understanding capabilities. Specifically, we provided serialized numerical coordinates in scatter plots and line graphs, defining a new coordinate space. Our trained model can accurately identify the localization of points in this coordinate space. As shown in Figure 5c is the document image with a scatter chart, and Figure 5f is the model's prediction, we only provided the vertical coordinates of the points in the image. However, the model can accurately identify their corresponding horizontal coordinates. For example, for a point with a vertical coordinate of 435.18, the model can identify its horizontal coordinate as 1096, which closely aligns with our provided ground truth.

4.3.2 Robust Interference Capability. Benefiting from training our model with documents containing natural images, our model exhibits robust interference capability. As shown in Figure 6, Figure 6a and Figure 6c presents a real image captured by a camera, while Figure 6b and Figure 6d illustrates the model's prediction. Despite incorrectly identifying some challenging regions as natural images, it does not impede subsequent text parsing. This phenomenon has not been observed in other end-to-end methods. We believe that training with synthetic data incorporating various contexts is an important approach to improving model robustness and performance.

5 Limitation

While the current generation of documents through SynthDoc is a significant step forward, we acknowledge that the types of documents created thus far are somewhat limited in variety. To enhance the richness of our dataset and to better mimic the complexity of real-world documents, we are committed to expanding our pipeline's capabilities. Future iterations will incorporate more sophisticated intermingling of document elements, allowing for the generation of even more intricate and varied document types. This evolution will not only challenge and refine existing models but also pave the way for the development of more advanced document image recognition systems, capable of handling the multifaceted nature of documents encountered in everyday applications.

6 Conclusion

In conclusion, this study presents SynthDoc, an innovative pipeline for generating synthetic documents, which plays a pivotal role in bolstering Visual Document Understanding (VDU). By producing a high-quality, diverse dataset that encompasses text, images, tables, and charts, SynthDoc addresses the critical issues of data acquisition and the constraints imposed by current datasets. Utilizing publicly accessible corpora and sophisticated rendering tools, SynthDoc has successfully created a dataset that is both extensive and adaptable. Our empirical evaluations, employing the Donut model, have shown that models trained on SynthDoc's dataset not only excel in pre-training read tasks but also exhibit resilience in downstream tasks, even when faced with linguistic disparities. The introduction of a benchmark dataset featuring 5,000 image-text pairs not only highlights the capabilities of our pipeline but also serves as a substantial contribution to the VDU community, facilitating further research and development in the realm of document image recognition. This research marks a significant advancement in the field by providing a scalable approach to overcoming data scarcity and by empirically validating the effectiveness of end-to-end models in parsing intricate, real-world documents.

References

[1] Srikar Appalaraju, Peng Tang, Qi Dong, Nishant Sankaran, Yichu Zhou, and R. Manmatha. 2023. DocFormerv2: Local Features for Document Understanding. arXiv:2306.01733 [cs.CV] https://arxiv.org/abs/2306.01733

[2] Haoli Bai, Zhiguang Liu, Xiaojun Meng, Wentao Li, Shuang Liu, Nian Xie, Rongfu Zheng, Liangwei Wang, Lu Hou, Jiansheng Wei, Xin Jiang, and Qun Liu. 2022. Wukong-Reader: Multi-modal Pre-training for Fine-grained Visual Document Understanding. arXiv:2212.09621 [cs.CL] https://arxiv.org/abs/2212.09621

[3] Sanket Biswas, Pau Riba, Josep Lladós, and Umapada Pal. 2021. DocSynth: a layout guided approach for controllable document image synthesis. In *International Conference on Document Analysis and Recognition*. 555–568.

[4] Ali Furkan Biten, Ruben Tito, Lluis Gomez, Ernest Valveny, and Dimosthenis Karatzas. 2022. Ocr-idl: Ocr annotations for industry document library dataset. In *European Conference on Computer Vision*. 241–252.

[5] Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents. arXiv:2308.13418 [cs.LG] https://arxiv.org/abs/2308.13418

[6] Brian Davis, Bryan Morse, Bryan Price, Chris Tensmeyer, Curtis Wigington, and Vlad Morariu. 2022. End-to-end document recognition and understanding with dessurt. In *European Conference on Computer Vision*. 280–296.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929 [cs.CV] https://arxiv.org/abs/2010.11929

[8] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. 2024. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems* 36 (2024), 27092–27112.

[9] Zhaojun Guo, Jinghui Lu, Xuejing Liu, Rui Zhao, ZhenXing Qian, and Fei Tan. 2024. What Makes Good Few-shot Examples for Vision-Language Models? arXiv:2405.13532 [cs.CV] https://arxiv.org/abs/2405.13532

[10] Kamal Gupta, Justin Lazarow, Alessandro Achille, Larry S Davis, Vijay Mahadevan, and Abhinav Shrivastava. 2021. Layouttransformer: Layout generation and completion with self-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1004–1014.

[11] Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis. 2015. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. 991–995.

[12] Teakgyu Hong, Donghyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2022. Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 10767–10775.

[13] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4083–4091.

[14] John D. Hunter. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* (2007), 90–95. https://doi.org/10.1109/MCSE.2007.55

[15] Wonseok Hwang, Seonghyeon Kim, Minjoon Seo, Jinyeong Yim, Seunghyun Park, Sungrae Park, Junyeop Lee, Bado Lee, and Hwalsuk Lee. 2019. Post-OCR parsing: building simple and robust parser via BIO tagging. In *Workshop on Document Intelligence at NeurIPS 2019*.

[16] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, Vol. 2. 1–6.

[17] Geewook Kim1, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. OCR-free Document Understanding Transformer. In *European Conference on Computer Vision*. 498–517.

[18] Hugo Laurençon, Lucile Saulnier, Leo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2024. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems* 36 (2024).

[19] Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems* 35 (2022), 31809–31826.

[20] Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, and Ming-Wei Chang nad Kristina Toutanova. 2023. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*. 18893–18912.

[21] David Lewis, Gady Agam, Shlomo Argamon, Ophir Frieder, David Grossman, and Jefferson Heard. 2006. Building a test collection for complex document information processing. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 665–666.

[22] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv:1910.13461 [cs.CL] https://arxiv.org/abs/1910.13461

[23] Deqing Li, Honghui Mei, Yi Shen, Shuang Su, Wenli Zhang, Junting Wang, Ming Zu, and Wei Chen. 2018. ECharts: a declarative framework for rapid construction of web-based visualization. *Visual Informatics* (2018), 136–146.

[24] Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. 2020. Tablebank: A benchmark dataset for table detection and recognition. arXiv:1903.01949 [cs.CV] https://arxiv.org/abs/1903.01949

[25] Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. 2020. DocBank: A benchmark dataset for document layout analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*. 949–960.

[26] Haofu Liao, Aruni RoyChowdhury, Weijian Li, Ankan Bansal, Yuting Zhang, Zhuowen Tu, Ravi Kumar Satzoda, R. Manmatha, and Vijay Mahadevan. 2023. Doctr: Document transformer for structured information extraction in documents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 19584–19594.

[27] Meng Ling, Jian Chen, Torsten Möller, Petra Isenberg, Tobias Isenberg, Michael Sedlmair, Robert S Laramee, Han-Wei Shen, Jian Wu, and C Lee Giles. 2021. Document domain randomization for deep learning document layout extraction. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16*. 497–513.

[28] Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Martin Eisenschlos. 2023. MatCha: Enhancing Visual Language Pretraining with Math Reasoning and Chart Derendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 12756–12770.

[29] Xuejing Liu, Wei Tang, Jinghui Lu, Rui Zhao, Zhaojun Guo, and Fei Tan. 2023. Deeply coupled cross-modal prompt learning. In *Findings of the Association for Computational Linguistics: ACL 2023*. 7957–7970.

[30] Xuejing Liu, Wei Tang, Xinzhe Ni, Jinghui Lu, Rui Zhao, Zechao Li, and Fei Tan. 2023. What Large Language Models Bring to Text-rich VQA? arXiv:2311.07306 [cs.CV] https://arxiv.org/abs/2311.07306

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10012–10022.

[32] Tengchao Lv, Yupan Huang, Jingye Chen, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, Li Dong, Weiyao Luo, Shaoxiang Wu, Guoxin Wang, Cha Zhang, and Furu Wei. 2024. Kosmos-2.5: A multimodal literate model. arXiv:2309.11419 [cs.CL] https://arxiv.org/abs/2309.11419

[33] Dhouib Mohamed, Bettaieb Ghassen, and Shabou Aymen. 2023. DocParser: End-to-end OCR-free Information Extraction from Visually Rich Documents. In *International Conference on Document Analysis and Recognition*. 155–172.

[34] The pandas development team. 2020. *pandas-dev/pandas: Pandas*. https://doi.org/10.5281/zenodo.3509134

[35] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, and Minjoon Seo. 2019. CORD: a consolidated receipt dataset for post-OCR parsing. In *Workshop on Document Intelligence at NeurIPS 2019*.

[36] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data only. *Advances in Neural Information Processing Systems* 36 (2023), 79155–79172.

[37] Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S Nassar, and Peter Staar. 2022. Doclaynet: A large human-annotated dataset for document-layout segmentation. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*. 3743–3751.

[38] Lorenzo Pisaneschi and et al.. 2023. Automatic generation of scientific papers for data augmentation in document layout analysis. *Pattern Recognition Letters* 167 (2023), 38–44.

[39] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* 35 (2022), 25278–25294.

[40] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv:2111.02114 [cs.CV] https://arxiv.org/abs/2111.02114

[41] Wei Tang, Liang Li, Xuejing Liu, Lu Jin, Jinhui Tang, and Zechao Li. 2024. Context disentangling and prototype inheriting for robust visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 5 (2024), 3213–3229.

[42] Zilong Wang, Yiheng Xu, Lei Cui, Jingbo Shang, and Furu Wei. 2021. Layoutreader: Pre-training of text and layout for reading order detection. arXiv:2108.11591 [cs.CL] https://arxiv.org/abs/2108.11591

[43] Maurice Weber, Carlo Siebenschuh, Rory Butler, Anton Alexandrov, Valdemar Thanner, Georgios Tsolakis, Haris Jabbar, Ian Foster, Bo Li, Rick Stevens, and Ce Zhang. 2024. WordScape: a Pipeline to extract multilingual, visually rich Documents with Layout Annotations from Web Crawl Data. *Advances in Neural Information Processing Systems* 36 (2024).

[44] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2023. Vary: Scaling up the vision vocabulary for large vision-language models. arXiv:2312.06109 [cs.CV] https://arxiv.org/abs/2312.06109

[45] Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü, Rui Hu, Chenxia Li, Liu Yang, Xilin Luo, Xuejie Wu, Lunan Liu, Wenjun Cheng, Peng Cheng, Jianhao Zhang, Xiaoyu Zhang, Lei Lin, Xiaokun Wang, Yutuan Ma, Chuanhai Dong, Yanqi Sun, Yifu Chen, Yongyi Peng, Xiaojuan Liang, Shuicheng Yan, Han Fang, and Yahui Zhou. 2023. Skywork: A More Open Bilingual Foundation Model. arXiv:2310.19341 [cs.CL] https://arxiv.org/abs/2310.19341

[46] WenmuZhou and SWHL. [n. d.]. TableGeneration. GitHub repository. https://github.com/WenmuZhou/TableGeneration

[47] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 1192–1200.

[48] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2022. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. arXiv:2012.14740 [cs.CL] https://arxiv.org/abs/2012.14740

[49] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. arXiv:2010.11934 [cs.CL] https://arxiv.org/abs/2010.11934

[50] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, Qin Jin, Liang He, Xin Lin, , and Fei Huang. 2023. UReader: Universal OCR-free Visually-situated Language Understanding with Multimodal Large Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 2841–2858.

[51] Xinru Zheng, Xiaotian Qiao, Ying Cao, and Rynson WH Lau. 2019. Content-aware generative modeling of graphic design layouts. *ACM Transactions on Graphics (TOG)* 38 (2019), 1–15.

[52] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. 2019. PubLayNet: Largest Dataset Ever for Document Layout Analysis. In *2019 International conference on document analysis and recognition (ICDAR)*. 1015–1022. https://doi.org/10.1109/ICDAR.2019.00166

Multimodal Understanding: Investigating the Capabilities of Large Multimodal Models for Object Detection in XR Applications

Rahel Arnold
University of Basel
Basel, Switzerland
rahel.arnold@unibas.ch

Heiko Schuldt
University of Basel
Basel, Switzerland
heiko.schuldt@unibas.ch

Abstract

Extended Reality (XR), encompassing the concepts of augmented, virtual, and mixed reality, has the potential to offer unprecedented types of user interactions. An essential requirement is the automated understanding of a user's current scene, for instance, in order to provide information via visual overlays, to interact with a user based on conversational interfaces, to provide visual clues on directions, to explain the current scene or even to use the current scene or parts thereof in automated queries. Key to scene understanding and thus to all these user interactions is high quality object detection based on multimodal content – images, videos, audio, etc. Large Multimodal Models (LMMs) seamlessly process text in conjunction with such multimodal content. Therefore, they are an excellent basis for novel XR-based user interactions, given that they provide the necessary detection quality.

This paper presents a two-stage analysis: In the first stage, the quality of two of the most prominent LMMs (LLaVA and KOSMOS-2) is compared with the object detector YOLO. The second step exploits Fooocus, a free and open-source AI image generator based on Stable Diffusion for the generation of images based on the descriptions derived in the first step. The second step evaluates the quality of the scene descriptions obtained in stage one. The evaluation results show that LLaVA, KOSMOS-2 and YOLO can all outperform the other approaches depending on the specific research focus. LLaVA achieves the highest recall, KOSMOS-2 results are the best in precision, and YOLO performs much faster and leads with the best F1 score. Fooocus manages to create images containing specific objects while still taking its liberty to omit or add specific objects. Therefore, our study confirmed our hypothesis that LMMs can be integrated into XR-based systems to further research novel XR-based user interactions.

CCS Concepts

• **Information systems → Query suggestion**; • **Computing methodologies → Machine learning**.

LGM3A '24, October 28-November 1, 2024, Melbourne, VIC, Australia
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1193-0/24/10
https://doi.org/10.1145/3688866.3689126

Keywords

Object Detection, LLM, LMM, Multimedia Retrieval, Extended Reality

ACM Reference Format:
Rahel Arnold and Heiko Schuldt. 2024. Multimodal Understanding: Investigating the Capabilities of Large Multimodal Models for Object Detection in XR 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, 10 pages. https://doi.org/10.1145/3688866.3689126

1 Introduction

The rapid advancements in *Large Language Models* (LLMs) [8] have enabled significant progress in natural language processing tasks, such as text classification, sentiment analysis, and machine translation. With over 180 million users, ChatGPT[1] by OpenAI is the most popular LLM. Other major companies like Meta, with the LLaMA model [44], provide LLMs that users can deploy locally. Unlike LLMs, *Large Multimodal Models (LMMs)* [51] possess the unique ability to process not only text but also multimodal content, including images, sound, and videos. This versatility extends their capabilities beyond text-based interactions. Recent studies [48, 49] have showcased the potential of LMMs in multimodal understanding, leveraging both linguistic and visual cues to identify objects, scenes, and actions.

LMMs play a crucial role in *eXtended Reality (XR)* applications, where users equipped with dedicated goggles or mobile devices perceive a real-world scene and get additional information on the objects within it. For these applications, high-quality LMMs for multimodal scene understanding are vital.

The potential of LMMs for XR applications in indoor architectures, a pivotal aspect of modern life, is profound. As buildings become more intricate and dynamic, indoor environments become more diverse and pose challenges to on-scene understanding to support XR applications. In such XR applications, LMMs could revolutionise how users navigate these varied indoor spaces, such as open-plan offices, museums, shopping malls, and healthcare facilities, each with unique navigation strategies.

Computer vision approaches have traditionally been employed to recognise objects and scenes within these environments. However, these methods rely on extensive training data, domain-specific expertise, and manual annotation. Moreover, these approaches may need help with complex or dynamic scenarios, such as varying lighting conditions, occlusions, or changes in the environment's

[1] https://chatgpt.com

layout. Additionally, they only recognise objects in isolation rather than understanding their relationships.

Multimodal understanding can play a vital role in indoor architecture to overcome these limitations. By combining linguistic and visual cues, users can interact with physical spaces using natural language queries, such as "Find the nearest coffee shop" or "Show me where the meeting room is located." This intuitive and user-friendly interaction style can significantly improve the user experience, reducing cognitive load and increasing efficiency – not exclusively, but very prominently in XR settings. Furthermore, multimodal understanding can enable more accurate and efficient object detection. For instance, a user might ask, "What is that object on the table?" or "Where can I get a similar object to the one standing there?" By analysing both linguistic cues (e.g., the object's description) and visual cues (e.g., the object's appearance), an LMM can accurately identify the object and provide relevant information.

This information needed by users can be translated into queries in multimedia retrieval. The concept would exist in XR to enable a user-interactive model where the digital and real worlds blend. Similar digital content can be retrieved by automatically detecting objects in the real world using an LMM. In the context of indoor architecture, users can retrieve the most similar objects in their collection and digitally place them (e.g., retrieved furniture). This potential application in mixed-reality multimedia retrieval opens up exciting possibilities for the future.

In this paper, we investigate the capabilities of Large Multimodal Models for object detection in images, exploring their potential to complement and surpass traditional computer vision approaches. We compare the performance of two LMMs (LLaVA [28, 29] and KOSMOS-2 [21, 21, 35]) with YOLO (You Only Look Once) [39, 47], a state-of-the-art object detector, on an indoor scene dataset. We aim to determine whether LMMs can recognise objects as effectively as, or better than, traditional object detection algorithms. In the second step, we explore image generation using Fooocus[2], an MLL capable of generating images. We investigate if a created image specified by the detected objects of an LMM contains the elements again by analysing the newly generated image with YOLO to use such automatically generated objects as query objects for a subsequent multimedia similarity search.

Our evaluation reveals that LLaVA excels in recall, making it particularly effective when identifying as many relevant objects as possible, which is critical. KOSMOS-2 demonstrates superior precision, indicating its strength in accurately identifying relevant objects with minimal false positives. On the other hand, YOLO is the fastest and achieves the highest F1 score, balancing precision and recall effectively. Furthermore, Fooocus, while creatively generating images with specified objects, sometimes introduces or omits elements, showing both the potential and limitations of image generation in XR applications. These findings confirm the potential of integrating LMMs into XR-based systems for enhanced user interactions and pave the way for future research in this domain.

The remainder of this paper is organised as follows: First, we review the foundations (Section 2) and present our method (Section 3). Then, we present our experiment and evaluate and discuss our findings (Section 4). Next, in Section 5 we provides an overview of related work. Lastly, Section 6 concludes and provides an outlook on future research and possible next steps.

2 Foundations

This chapter introduces an overview of computer vision before delving deeper into object detection, focusing on YOLO. We then explore the capabilities of *Multimodal Understanding* and *Large Multimodal Models*, highlighting their applications in multimedia retrieval and *eXtended Reality*.

2.1 General Computer Vision

Computer vision [4] is a field of artificial intelligence that helps computers understand visual data. It allows computers to interpret visual information and make decisions based on what they "see" in the world. This field involves creating algorithms and models to automate tasks our human visual system can perform. It has made significant progress in various fields, such as autonomous vehicles [23] and medical imaging [16]. These tasks include image processing [34], feature extraction [31], object detection [17], and image recognition [20, 22].

2.2 Object Detection

Object detection [17] is a crucial area within computer vision that involves identifying and localising objects within an image. Unlike image classification [26], which only assigns a label to an image, object detection provides the class and the bounding box coordinates for each object detected in the image. This makes it particularly useful for applications requiring precise object localisation, such as autonomous driving, surveillance, and extended reality.

Identifying and classifying objects are pivotal aspects of computer vision, enabling the development of applications ranging from autonomous robots to virtual assistants and smart home systems. Established object detection algorithms such as YOLO (You Only Look Once) [39], SSD (Single Shot Detector) [30], and Faster R-CNN (Region-based Convolutional Neural Networks) [40] have demonstrated notable achievements across various domains.

YOLO [39] is a state-of-the-art object detection algorithm known for its speed and accuracy. Unlike traditional methods that apply a model to an image at multiple locations and scales, YOLO reframes object detection as a single regression problem, going straight from image pixels to bounding box coordinates and class probabilities. YOLO divides the image into a grid and predicts bounding boxes and probabilities for each grid cell. Its main advantages are its high speed and real-time processing capabilities, making it ideal for applications requiring rapid object detection. YOLO has undergone several iterations, with YOLOv10 [47] being the newest. These models leverage advancements in convolutional neural networks (CNNs) and feature extraction techniques to enhance object detection capabilities.

In contrast, SSD [30] is an expedited algorithm that utilises a single neural network to forecast object locations and classes. However, SSD may demonstrate reduced accuracy compared to YOLO in detecting small or partially occluded objects. Faster R-CNN, incorporating region proposal networks (RPNs) for object detection, has exhibited superior accuracy to traditional algorithms like YOLO and

[2] https://github.com/lllyasviel/Fooocus

SSD. Nonetheless, Faster R-CNN [40] may entail high computational demands and limited performance in real-time applications [25].

2.3 LLMs, MU, and LMMs

Large Language Models (LLMs) [8] are advanced AI-powered systems that understand and generate human language. They use deep learning techniques and extensive datasets to revolutionise natural language processing (NLP) [12]. LLMs enable various language-related tasks such as translation, summarisation, and conversational agents. An example of LLMs is GPT-3.5 [8], which OpenAI developed. Another prominent LLM is LLaMA (Large Language Model Meta AI) [44], created by Meta, which prioritises efficiency and scalability, making it a valuable tool for both research and practical NLP applications. These models signify significant advancements in AI, providing robust solutions for complex linguistic challenges.

Multimodal understanding (MU) encompasses processing multiple data modalities, such as images, videos, text, and audio, to comprehend the scene entirely [10].

Large Multimodal Models (LMMs) [51] are extensions of LLMs that integrate these multiple data inputs into a single framework. Neural networks build their architectures. This multimodal approach enables a more holistic understanding of data, allowing models to leverage different types of information simultaneously. LMMs are trained on vast datasets that include paired text and image data, enabling them to learn relationships between visual and linguistic information. This capability allows LMMs to perform tasks that require understanding text and visual content, such as image captioning, visual question answering, and multimodal search. Models of LMMs are, for example, LLaVA (Large Language and Vision Assistant) [28, 29], which integrates visual and textual data to perform tasks like visual question answering and image captioning, KOSMOS-2 [21, 21, 35], developed by Microsoft, which can seamlessly integrate textual descriptions into corresponding bounding boxes within images, enhancing the model's ability to understand and generate contextually accurate visual and textual data, or GPT-4 [32], created by OpenAI. Other LMMs are not capable of analysing but generating new content. DALL-E, operated by OpenAI, Midjourney, and Fooocus, are examples of such LMMs [6].

2.4 Multimedia Retrieval

Multimedia retrieval involves searching and retrieving relevant multimedia content (such as images, videos, and audio) based on a query [41]. Traditional multimedia retrieval systems rely heavily on metadata and manually annotated tags. However, advancements in computer vision and multimodal learning have enabled more sophisticated retrieval systems to interpret and understand the content within multimedia data itself [18, 19, 45]. Multimodal models enhance multimedia retrieval by allowing users to search using natural language queries describing visual content. This approach can significantly improve the accuracy and relevance of search results, making it easier for users to find the information they need. CLIP (Contrastive Language–Image Pre-training) [36] is one of the most used features nowadays. Its architecture employs a vision and text encoder jointly trained to encode images and their textual descriptions into the same space.

2.5 Extended Reality

Extended reality (XR) [14, 53] is an umbrella term for all immersive technologies *Augmented Reality (AR)*, *Mixed Reality (MR)*, and *Virtual Reality (VR)*. Depending on the exact technology, environments with physical and virtual objects exist and interact with each other.

Integrating LMMs in XR can revolutionise how users interact with their environments. By leveraging multimodal understanding, XR systems can recognise and interpret objects and scenes in the real world, enabling more natural and intuitive interactions. For example, an XR system could recognise furniture in a room and provide digital overlays or suggestions for rearranging the space. This capability has significant potential in various applications, including architecture, interior design, and urban planning.

3 Methodology

This chapter outlines the methodology employed to compare the performance of YOLO with two LMMs, LLaVA and KOSMOS-2, for object detection in indoor architectural environments. The primary focus is to evaluate the effectiveness of these models in recognising and identifying objects within images, which is essential for applications in mixed-reality multimedia retrieval. In the second step, we consider the image generation of an LMM (Fooocus) by specifying objects that should be present in the generated file.

3.1 Experimental Design

We evaluate the object detection performance of a specified state-of-the-art algorithm compared to asking two LMMs to describe the objects visible in an image and one LMM for multimodal generation. Therefore, we assess how well YOLO, LLaVA, and KOSMOS-2 detect and identify objects in a curated dataset of indoor scenes and how precisely Fooocus generates images by specifying visible objects.

3.1.1 Dataset Preparation. To conduct the object detection evaluation, we use a comprehensive dataset featuring various indoor environments [1]. It contains 4, 147 images from interior designs, such as living rooms, bedrooms, and kitchens. The models process each image in the dataset. We selected a random subset of 100 pictures and annotated the detected objects manually. This annotation is considered as the ground truth. We consciously chose this data set because it covers indoor architecture. As our work is situated in this context, it is a perfect fit. An alternative option would have been Microsoft COCO: Common Objects in Context [11], which is already annotated and, therefore, has a ground truth. However, as the dataset does not concern interior architecture, we decided not to use it.

3.1.2 Model Selection and Configuration.

YOLO (YOLOv10-M) [47]: The latest iteration of YOLO, pre-trained on the COCO dataset [27]. YOLOv10 is chosen for its balance between speed and accuracy, making it suitable for real-time object detection tasks. In total, it detects 80 different classes of objects.

LLaVA (llava-hf/llava-v1.6-mistral-7b-hf) [28]: This large multimodal model integrates visual and textual data. Leveraging its pre-training on extensive paired text-image datasets, LLaVA is configured to process the images in our dataset. Its ability to understand the context from both visual and

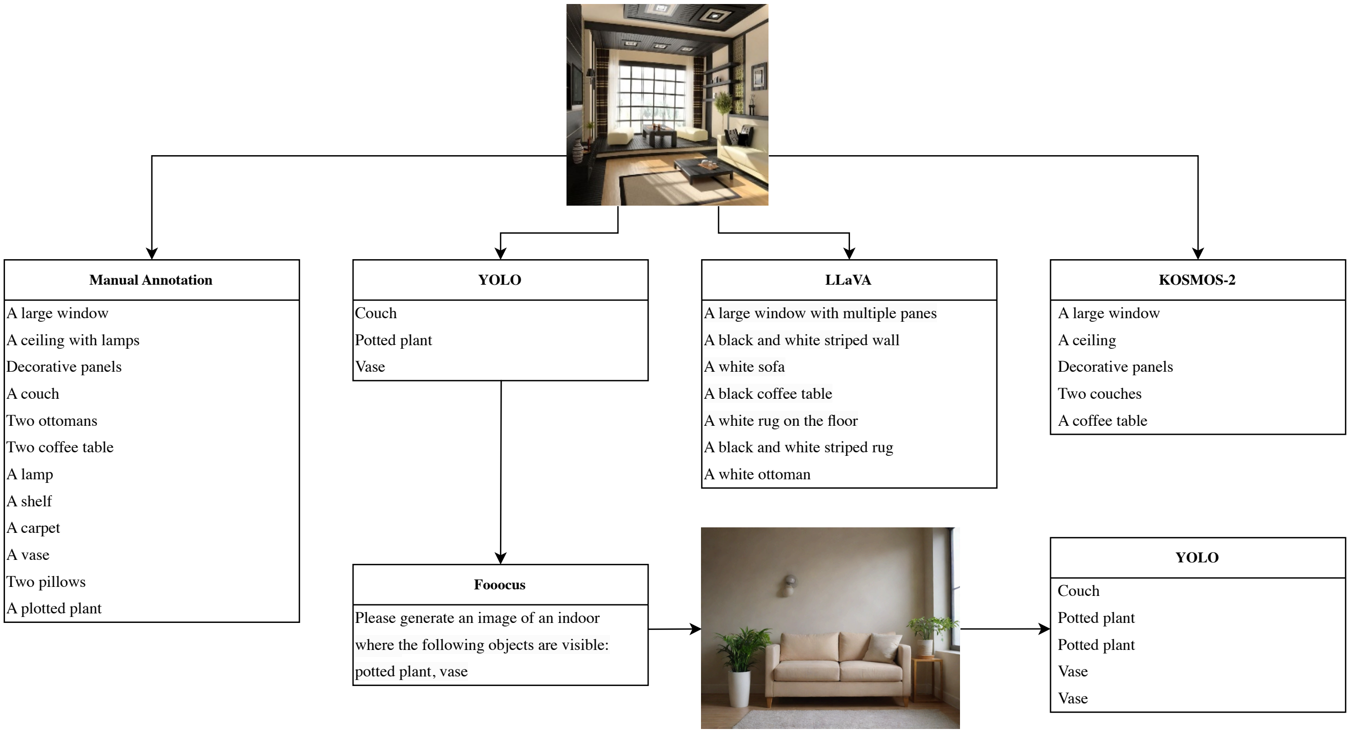


Figure 1: This figure presents the object detection results of YOLO, LLaVA, and KOSMOS-2 on an arbitrary image from the indoor environments dataset. The different models do not recognise the same objects. As visible, the output of YOLO is further processed by Fooocus to generate images, where we apply YOLO again.

textual cues is crucial for effective object detection without the limitation of predefined classes.

KOSMOS-2 (microsoft/kosmos-2-patch14-224) [35]: Developed by Microsoft, KOSMOS-2 is another LMM fine-tuned specifically for object detection within our dataset. Its architecture is optimised for interpreting and integrating multimodal data, making it a valuable model for our comparative study. Again, no limitation regarding classified objects exists.

Fooocus ("juggernautXL_v8Rundiffusion"): The default model of Fooocus, a free and open-source AI image generator based on *Stable Diffusion*, is employed to generate images [15].

3.1.3 Evaluation Metrics. To measure the performance of the models, we employ several evaluation metrics:

Precision and Recall: Precision measures the percentage of correctly identified objects, while recall calculates the percentage of actual objects correctly identified. These metrics provide insights into the accuracy and comprehensiveness of each model's object detection capabilities.

F1 Score: The F1 score is the harmonic mean of precision and recall, offering a single metric that balances the trade-off between these two measures.

Inf. Time: Inference time refers to each model's time to process and detect objects in an image. This metric is crucial for real-time object detection applications like extended reality environments.

Object Retention Rate (ORR): This metric measures how many objects from the original image are retained in the newly generated images by the Fooocus API. It is calculated as the number of original objects detected in the generated image divided by the total number of original objects.

New Object Introduction Rate (NOIR): NOIR evaluates how many new objects are detected in the generated images that were not part of the original image. It is calculated as the number of new objects detected in the generated image divided by the total number of detected objects in the generated image.

3.1.4 Procedure. The procedure for the object detection evaluation is detailed as follows.

(1) Image Processing: Each image in the dataset is processed by YOLO, LLaVA, and KOSMOS-2. The models analyse the images and generate labels for detected objects. The generated labels of YOLO are parsed as a prompt to Fooocus to generate a new image. YOLO is applied again. This pipeline is shown in Figure 1.
(2) Output Comparison: The outputs from each model are compared against the manually annotated ground truth. This comparison involves checking the labels for accuracy and calculating the evaluation metrics.
(3) Performance Calculation: Each model's precision, recall, F1 score, inference time, ORR and NOIR for image generation

Table 1: Performance Metrics for Object Detection

Model	Precision	Recall	F1 Score	Avg. Inf. Time
YOLO	0.57	0.38	0.43	12.09 ms
LLaVA	0.30	0.58	0.35	3719.41 ms
KOSMOS-2	0.65	0.35	0.42	1492.79 ms

are calculated and recorded. These metrics are used to assess and compare the performance of YOLO, LLaVA, and KOSMOS-2 in detecting objects within indoor scenes and to evaluate Fooocus generating images with object requirements.

3.2 Application in Extended Reality

The findings from our study have significant implications for XR applications. By understanding how well LMMs and traditional object detection models perform in identifying objects within indoor environments, we can enhance the interactive capabilities of extended-reality systems, as in [3] with an MR system. Accurate detection and contextual understanding of objects allow for seamless integration of digital content into the real world, enabling applications such as virtual interior design, where users can visualise and interact with virtual furniture and decor within their physical spaces. XR systems can provide more accurate and context-aware guidance within complex indoor environments, and museums and educational facilities can leverage mixed reality to provide immersive and informative experiences by overlaying digital information on physical exhibits.

4 Experiments, Results and Discussion

This section presents the experiments conducted to evaluate the performance of YOLO, LLaVA, KOSMOS-2, and Fooocus in object detection and image generation. We follow the methodology presented in the previous section. We illustrate and discuss the results obtained.

4.1 Experimental Setup

We conducted experiments using a curated dataset of indoor environments containing 4, 147 images. YOLO, LLaVA, and KOSMOS-2 processed each image to detect and identify objects. Additionally, the detected objects from YOLO were used as prompts to generate new images with Fooocus, and the generated images were further analysed using YOLO.

We performed the evaluation on a server configured with the following specifications: The operating system used was Ubuntu 22.04.4 LTS, running on a Linux kernel version 5.15.0-113-generic, optimised for x86_64 architecture. The server is equipped with an NVIDIA GeForce RTX 3070 graphics card. The driver version installed was 535.183.01, which supports the CUDA 12.2 toolkit. The CUDA version is critical to ensure compatibility with GPU-accelerated computation's latest features and optimisations.

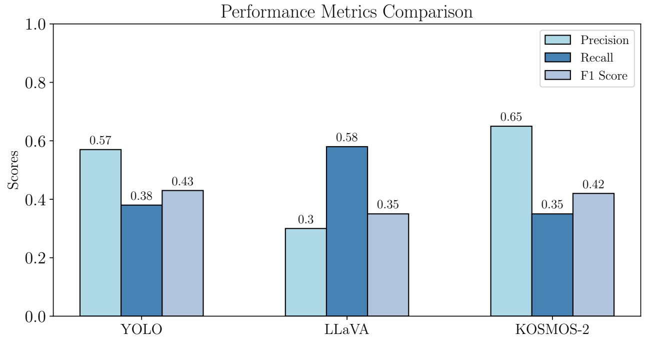


Figure 2: Comparison of performance metrics for object detection models. The bar plot displays the precision, recall, and F1 scores for YOLO, LLaVA, and KOSMOS-2 models.

4.2 Results Processing

After completion of all the evaluations, we carry out model-specific post-parsing processes. This step is necessary because not all models return results in the same format.

From YOLO, we receive the label of the recognised object, the confidence and the bounding box. However, only the label is essential for our evaluation. With KOSMOS-2, the results of the individual detections look similar to the YOLO results as we receive the object and its position, i.e., the bounding box, in the image again.

With LLaVA, however, the results are less structured as they come as free text. Sometimes, the response consists of one sentence containing a long enumeration, but other times, it consists of several sentences. To structure the queries as uniformly as possible, the user prompted the LLaVA command to name the recognized objects in a list. Sometimes, the response starts with phrases such as "Sure, here are the objects I see in the picture...".

Consequently, with this model, we receive many words that have nothing to do with the recognised objects. Therefore, our first preprocessing step is to filter out all nouns with the help of the NLTK corpus[3], as only these can stand for objects.

The further procedure was identical for all models. Of the 100 randomly selected images, we checked which objects were also recognised by the object detection model. However, since an object can also be described differently using synonyms, we use synsets[4] to check whether the recognised objects are synonyms.

4.3 Results

Our study's results are divided into two categories for analysis: Object Detection Performance and Image Generation. This approach allows for a comprehensive examination of both the efficacy of object detection and the quality of generated images.

4.3.1 Object Detection Performance. The object detection performance of YOLO, LLaVA, and KOSMOS-2 was evaluated based on precision, recall, F1 score, and inference time. The exact results of our evaluation are presented in Table 1.

[3] https://www.nltk.org/api/nltk.corpus.html
[4] https://www.nltk.org/howto/wordnet.html

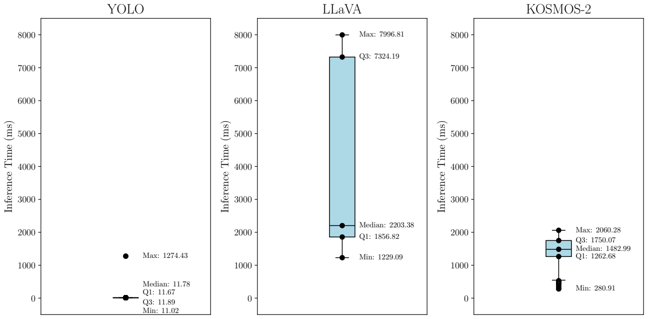


Figure 3: Inference Time Distribution for YOLO, LLaVA, and KOSMOS-2 Object Detection

Figure 2 graphically illustrates each model’s precision, recall, and F1 score results. YOLO has high precision but low recall, while LLaVA demonstrates the best recall performance but lower precision and F1 score. KOSMOS-2 has the highest precision but the lowest recall. Therefore, YOLO is considered the most balanced, as it has the highest F1 score, even though it is not exceptionally high.

Looking at the times the models used for object recognition, we can see that YOLO can achieve the fastest times by far. With an average inference time of about 12 ms, it is almost 125 times as fast as KOSMOS and over 300 times as fast as LLaVA, presented in Table 1. It took YOLO for over 75% of all images less than 12 ms to detect the objects, as can be seen in Figur 3. For KOSMOS-2, the times are in the interquartile range between 1.2 and 1.7 seconds. On average, the determination takes around 1.5 seconds. With LLaVA, the inferences are significantly different magnitudes, with the media still at 2.2 seconds and the average at 3.7 seconds. Some analyses take over 7 seconds. Consequently, this model is by far the slowest.

4.3.2 Image Generation and Evaluation. Fooocus was used to generate new images based on object detections from YOLO. We measure the quality of these generated images by measuring the Object Retention Rate (ORR) and the New Object Introduction Rate (NOIR).

When we examine the ORR and NOIR data, we find that, on average, 50% of the desired objects reappeared in the generated image. On average, slightly over 30% of new objects were added. This behaviour is also illustrated in Figure 4. The interquartile range shows higher values for existing objects compared to new ones. We also see that in some cases, all objects were correctly depicted, while in other instances, no objects remained. New objects did not always appear, but after classification with YOLO, some images consisted entirely of completely new objects.

Table 2: Performance Metrics for Image Generation with Fooocus

	ORR	NOIR
Average	0.57	0.35
Median	0.50	0.33
Standard Deviation	0.27	0.27

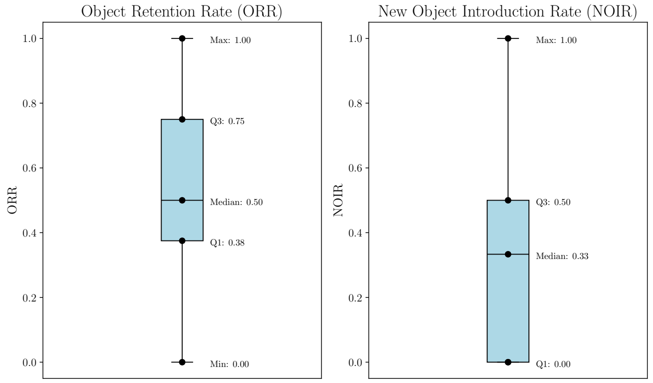


Figure 4: The left plot shows the distribution of ORRs, while the right plot presents the distribution of NOIR. The higher ORR median indicates maintaining the original objects, while the lower NOIR represents fewer newly generated objects.

The relationship between ORR and NOIR is captured in Figure 5. When ORR is high and the NOIR is low, the generated images retain most original objects and do not introduce many new ones. This outcome is desirable if the goal is to preserve the original content and avoid introducing new elements, as in our research. Low ORR and high NOIR indicate that the generated images do not retain many original objects but introduce many new ones. Therefore, the result is a loss of original content and a focus on adding new elements. High ORR and high NOIR show that the generated images retain most original objects but also introduce many new ones. This represents a balance between retaining original content and introducing new elements. Low ORR and low NOIR indicate that the generated images do not retain many original objects and also do not introduce many new objects. Consequently, the generated images are neutral and not very informative.

As we can see, our data points are very distributed. In image generation with Fooocus, objects are not always present as expected. The same applies to adding additional objects. However, based on the trend line, we notice that the more objects correctly inserted into the image, the fewer new objects are added. Fewer objects were created in images in which the desired objects were included than when Fooocus did not employ the desired objects.

4.4 Discussion

When we discuss and analyse the results, we first note that LMMs can recognise objects in images. However, it is also clear that neither the dedicated object detection algorithm YOLO nor the models LLaVA or KOSMOS-2 can entirely and accurately classify objects. It is important to note that the ground truth for the 100 sampled images was manually created. While efforts were made to ensure accuracy, minor errors or omissions may exist. However, these are not expected to lead to significant deviations in the overall analysis.

YOLO generally performs well. However, it should be noted that YOLO only recognises objects from predefined classes. Consequently, some objects are not considered because they are unknown to the model. This is likely an explanation for the lower recall value.

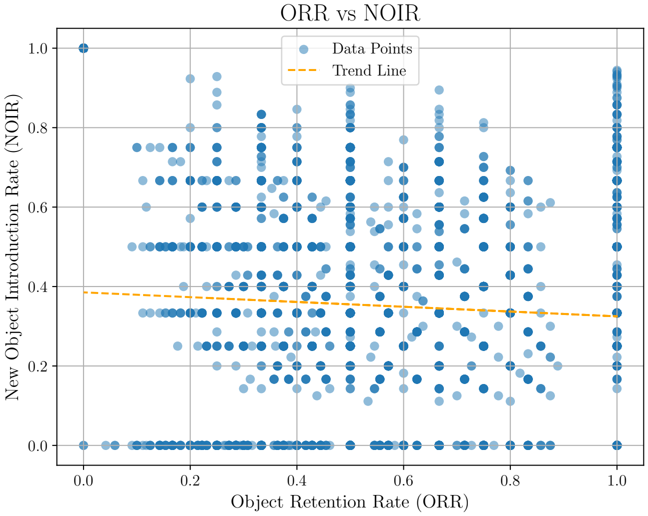


Figure 5: This scatter plot of ORR (x-axis) versus NOIR (y-axis) provides insights into the relationship between how well objects from the original images are retained in the generated images and how many new objects are introduced. Both axes are scaled from 0 to 1. The slightly negative slope of the trend line reveals that as the ORR increases, the NOIR tends to decrease slightly.

Objects are mostly recognised if they exist, reflected in the higher precision value. Additionally, the inference time favours YOLO. With its fast evaluation, the algorithm suits real-time applications in extended reality.

The high recall of LLaVA is undoubtedly very positive and encouraging, indicating that many correct objects were detected. At the same time, the low precision value must be mentioned. On one hand, this is likely due to receiving a full-text response. Even when we only check the nouns, some words do not correspond to detected objects. On the other hand, sporadic sample views of the detected elements also showed that the LMM occasionally invented new objects, leading to a low precision value. LLaVA's time to recognise objects is relatively high compared to other models. Therefore, this variant is only suitable for applications that are not too time-critical. However, since LLaVA provides both the object and a full description, it remains an exciting model for XR multimedia retrieval applications. Users can focus on details with LLaVA, which was impossible with YOLO until now.

KOSMOS-2 presents a comparable candidate to YOLO, with improved precision. The advantage is that it can still identify precise object boundaries in the image and is not limited to predefined classes. However, this model requires more time, though it is still within a reasonable range. In direct comparison to YOLO, a choice can definitely be made between speed and precision.

When we discuss how well LMMs create images based on object specifications, we find that they deliver solid performance, with more than half of the desired objects typically present. Of course, there can still be deviations, as YOLO does not always correctly classify all objects. This image generation approach is undoubtedly promising for suggestions for further searches in multimedia retrieval. If the objects are correct but the setting is not, a new example image can be generated based on the desired format and objects, which can then be used for similarity searches.

In summary, LMMs have a valid use case and offer exciting options for multimedia retrieval in XR compared to classical object detection algorithms. YOLO, LLaVA, and KOSMOS-2 can all excel depending on the focus.

5 Related Work

The rapid development of LLMs has sparked significant interest in their potential applications beyond traditional text-based tasks. LLMs like GPT-3 and ChatGPT have demonstrated impressive capabilities in natural language processing, including tasks such as text classification, sentiment analysis, and machine translation. However, the emergence of LMMs marks a notable shift in the field, extending these capabilities to process and understand multimodal content, including images, sound, and video. In this section, we will provide an overview of related work focusing on LMM, a comparison to traditional computer vision algorithms and their use in extended reality.

5.1 Multimodal Understanding with LMMs

LLMs have shown promising results in bridging the semantic gap due to their impressive multimodal capabilities. Learning visual concepts from natural language descriptions, LMMs enable efficient zero-shot transfer to various vision tasks [36]. This approach underscores the potential of LMMs in recognising and generating content that aligns with visual and textual inputs, even though they have not been trained on the data before. In this study, we focused on available open-source models with local deployment. However, closed-source or only cloud-available models exist, such as DALL-E [5, 37, 38], Midjourney[5], Claude[6] or Gemini [43].

5.2 Comparing LMM Object Detection to Traditional Computer Vision Algorithms

Incorporating object detection in LMMs offers several advantages over traditional methods. LMMs can provide a more holistic understanding of scenes by combining linguistic and visual cues, allowing for natural language interactions and context-aware object recognition. For instance, [52] demonstrated that incorporating language and vision models improves understanding and interpreting complex scenes in visual question-answering tasks.

The integration of object detection with large language models aims to leverage visual and linguistic understanding to enhance performance on multimodal tasks. Recent research has made significant advancements in this field, particularly in the following key areas:

Contextual Object Detection: This technique integrates LLMs with object detection models to extract context-specific image information. For example, the ContextDET framework combines a visual encoder, a pre-trained LLM, and a visual decoder to perform tasks such as cloze tests, captioning, and question answering by analysing images and human text

[5] https://www.midjourney.com
[6] https://claude.ai

inputs concurrently, providing both bounding boxes and textual outputs [50].

Enhancing LMMs with Detection Models: Another method focuses on integrating cutting-edge object detection and OCR models into multimodal large language models. A study [24] has demonstrated that models like LLaVA-1.5 can be optimized with the DINO object detection model [9, 33] toto enhance object counting and localisation accuracy. This integration converts detection model outputs into textual format and complements the LLM, enabling the model to utilise detailed image information alongside overall image data.

These developments showcase the expanding possibilities and applications of merging object detection with LMMs, advancing multimodal understanding and interaction.

5.3 LLMs and LMMs in Extended Reality

Integrating LLMs in XR is a relatively new research area yet to be explored and opens much potential. This is especially true for LMMs in this domain. One current explored possibility is to employ LLMs as conversation agents in XR, as different work shows [2, 42, 46]. These approaches show the use of LLMs in XR but no essential modification to adapt to the new application domain.

A recent study discusses integrating LLMs into XR environments to enhance inclusion, engagement, and privacy [7]. The research highlights how fine-tuning LLMs and employing different prompting techniques can improve task performance and personalisation in XR settings. Additionally, the study explores using custom LLMs in XR applications as smart non-player characters (NPCs).

Another approach is to use LLMs to create and edit objects and scenes in MR, as [13] examined. They demonstrated cross-platform interoperability with several example worlds. Furthermore, they evaluated their approach to various creation and modification tasks to show how a combination of LLMs can produce and edit diverse objects, tools, and scenes in MR, leading to a positive user interaction.

6 Conclusions and Future Work

Our analysis demonstrated the potential of LMMs in enabling novel user interactions for XR applications, confirming our hypothesis. By comparing the performance of the object detection algorithm YOLO with the two LMMs, LLaVA and KOSMOS-2, on an image dataset representing indoor scenes, we showed that each model has strengths and weaknesses depending on the specific research focus.

The results of our study highlight the importance of considering each model's benefits and limitations when designing XR-based applications. For instance, LLaVA's high recall and ability to provide free-text answers make it an excellent choice for applications where accuracy and detailed user interaction are vital. On the other hand, KOSMOS-2's precision ensures that detected objects are accurate and reliable, which is crucial for applications that require high confidence in object identification. Meanwhile, YOLO's speed and F1 score make it a suitable option for real-time applications in extended reality, where quick processing and balanced accuracy are important.

Furthermore, our study demonstrates the potential of LMMs to generate high-quality scene descriptions that can be used as input for XR-based systems. We observed that Fooocus, an AI image generator based on Stable Diffusion, can create images containing specific objects while also having the flexibility to omit or add certain elements. This ability to generate and modify scene content is beneficial for creating dynamic and adaptable XR environments.

The implications of our study extend beyond theoretical insights and have practical significance for developing novel XR-based user interactions. By leveraging LMMs' strengths and capability to integrate multimodal content, we can create innovative solutions that enhance users' interactions with buildings and spaces. For example, architects and designers can use LMMs to generate detailed descriptions of building designs and environments, facilitating more accurate and efficient design processes. This can lead to improved visualization and planning in architectural projects.

Additionally, our study highlights the potential of LMMs to improve the user experience in XR-based applications. High-quality, accurate, and reliable scene descriptions provided by LMMs can enable users to understand better and navigate complex spaces. This has significant implications for various industries, including architecture and interior design, where understanding spatial relationships and details is crucial.

Future work involves incorporating LLaVA into an XR system, especially XR multimedia retrieval, to enable users to interact more effectively with detected objects. This would allow for more accurate and efficient search queries and improved navigation and exploration of complex spaces. The user interaction could benefit as well. Additionally, Fooocus could likewise be employed to generate possible images that can be used for further similarity searches, enabling users to refine their search results and explore related concepts. By combining the strengths of LMMs with advanced multimedia retrieval techniques, our research can potentially impact the development of innovative XR-based applications. It would furthermore be interesting to see how well our concept works in other environments. Therefore, a follow-up study with the COCO dataset [11] could be carried out to potentially expand the range of applications. Another potential work would deal with our presented method and 3D models. Since the result is to be used in XR applications, not only 2D entities but also 3D objects are of interest. 3D scene understanding would, therefore, be a core component.

In conclusion, our study demonstrates the potential of LMMs in enabling novel user interactions for XR-based applications and suggests a potential revolution in the field. By understanding the strengths and limitations of each model, we can develop innovative solutions that enhance how users interact with buildings and spaces. The results of our study have significant implications for various industries and demonstrate the potential of LMMs to revolutionize the field of XR.

Acknowledgments

This work was funded by the Horizon Europe project XR mEdia eCOsystem (XReco) through a grant from the Swiss State Secretariat for Education, Research and Innovation (SERI), contract number 22.00268.

References

[1] Sofea Aishah. 2020. interior_design. https://www.kaggle.com/datasets/aishahsofea/interior-design

[2] Siddhanth Jayaraj Ajri, Dat Nguyen, Swati Agarwal, Arun Kumar Reddy Padala, and Caglar Yildirim. 2023. Virtual AIVantage: Leveraging Large Language Models for Enhanced VR Interview Preparation among Underrepresented Professionals in Computing. In *Proceedings of the 22nd International Conference on Mobile and Ubiquitous Multimedia (MUM '23)*. Association for Computing Machinery, New York, NY, USA, 535–537. https://doi.org/10.1145/3626705.3631799

[3] Rahel Arnold and Heiko Schuldt. 2024. Multimedia Retrieval in Mixed Reality: Leveraging Live Queries for Immersive Experiences. In *2024 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR)*. 289–293. https://doi.org/10.1109/AIxVR59861.2024.00048 ISSN: 2771-7453.

[4] Dana Harry Ballard and Christopher M. Brown. 1982. *Computer Vision* (1st ed.). Prentice Hall Professional Technical Reference.

[5] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. [n. d.]. Improving Image Generation with Better Captions. ([n. d.]).

[6] Ali Borji. 2023. Generated Faces in the Wild: Quantitative Comparison of Stable Diffusion, Midjourney and DALL-E 2. https://doi.org/10.48550/arXiv.2210.00586 arXiv:2210.00586 [cs].

[7] Efe Bozkir, Süleyman Özdel, Ka Hei Carrie Lau, Mengdi Wang, Hong Gao, and Enkelejda Kasneci. 2024. Embedding Large Language Models into Extended Reality: Opportunities and Challenges for Inclusion, Engagement, and Privacy. In *ACM Conversational User Interfaces 2024*. 1–7. https://doi.org/10.1145/3640794.3665563 arXiv:2402.03907 [cs].

[8] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. https://doi.org/10.48550/arXiv.2005.14165 arXiv:2005.14165 [cs].

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging Properties in Self-Supervised Vision Transformers. https://doi.org/10.48550/arXiv.2104.14294 arXiv:2104.14294 [cs].

[10] Wei Chen, Weiping Wang, Li Liu, and Michael S. Lew. 2021. New Ideas and Trends in Deep Multimodal Content Understanding: A Review. *Neurocomputing* 426 (Feb. 2021), 195–215. https://doi.org/10.1016/j.neucom.2020.10.042

[11] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. https://doi.org/10.48550/arXiv.1504.00325 arXiv:1504.00325 [cs].

[12] K. R. Chowdhary. 2020. Natural Language Processing. In *Fundamentals of Artificial Intelligence*, K.R. Chowdhary (Ed.). Springer India, New Delhi, 603–649. https://doi.org/10.1007/978-81-322-3972-7_19

[13] Fernanda De La Torre, Cathy Mengying Fang, Han Huang, Andrzej Banburski-Fahey, Judith Amores Fernandez, and Jaron Lanier. 2024. LLMR: Real-time Prompting of Interactive Worlds using Large Language Models. http://arxiv.org/abs/2309.12276 arXiv:2309.12276 [cs].

[14] Lucio Tommaso De Paolis, Pasquale Arpaia, and Marco Sacco (Eds.). 2023. *Extended Reality: International Conference, XR Salento 2023, Lecce, Italy, September 6-9, 2023, Proceedings, Part II*. Lecture Notes in Computer Science, Vol. 14219. Springer Nature Switzerland, Cham. https://doi.org/10.1007/978-3-031-43404-4

[15] Nassim Dehouche and Kullathida Dehouche. 2023. What's in a text-to-image prompt? The potential of stable diffusion in visual arts education. *Heliyon* 9, 6 (June 2023), e16757. https://doi.org/10.1016/j.heliyon.2023.e16757

[16] Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. 2021. Deep learning-enabled medical computer vision. *npj Digital Medicine* 4, 1 (Jan. 2021), 1–9. https://doi.org/10.1038/s41746-020-00376-2 Publisher: Nature Publishing Group.

[17] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. 2010. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 9 (Sept. 2010), 1627–1645. https://doi.org/10.1109/TPAMI.2009.167 Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[18] Ralph Gasser, Rahel Arnold, Fynn Faber, Heiko Schuldt, Raphael Waltenspül, and Luca Rossetto. 2024. A New Retrieval Engine for Vitrivr. In *MultiMedia Modeling*, Stevan Rudinac, Alan Hanjalic, Cynthia Liem, Marcel Worring, Björn Þór Jónsson, Bei Liu, and Yoko Yamakata (Eds.). Springer Nature Switzerland, Cham, 324–331. https://doi.org/10.1007/978-3-031-53302-0_28

[19] Ralph Gasser, Luca Rossetto, and Heiko Schuldt. 2019. Multimodal Multimedia Retrieval with vitrivr. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*. ACM, Ottawa ON Canada, 391–394. https://doi.org/10.1145/3323873.3326921

[20] Ernest Hall. 1979. *Computer Image Processing and Recognition*. Elsevier.

[21] Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. 2022. Language Models are General-Purpose Interfaces. https://doi.org/10.48550/arXiv.2206.06336 arXiv:2206.06336 [cs].

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. https://doi.org/10.1109/CVPR.2016.90 ISSN: 1063-6919.

[23] Joel Janai, Fatma Güney, Aseem Behl, and Andreas Geiger. 2020. Computer Vision for Autonomous Vehicles: Problems, Datasets and State of the Art. *Foundations and Trends® in Computer Graphics and Vision* 12, 1–3 (July 2020), 1–308. https://doi.org/10.1561/0600000079 Publisher: Now Publishers, Inc..

[24] Qirui Jiao, Daoyuan Chen, Yilun Huang, Yaliang Li, and Ying Shen. 2024. Enhancing Multimodal Large Language Models with Vision Detection Models: An Empirical Study. https://doi.org/10.48550/arXiv.2401.17981 arXiv:2401.17981 [cs].

[25] Jeong-ah Kim, Ju-Yeong Sung, and Se-ho Park. 2020. Comparison of Faster-RCNN, YOLO, and SSD for Real-Time Vehicle Type Recognition. In *2020 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia)*. 1–4. https://doi.org/10.1109/ICCE-Asia49877.2020.9277040

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, Vol. 25. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft COCO: Common Objects in Context. https://doi.org/10.48550/arXiv.1405.0312 arXiv:1405.0312 [cs].

[28] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved Baselines with Visual Instruction Tuning. https://doi.org/10.48550/arXiv.2310.03744 arXiv:2310.03744 [cs].

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. https://doi.org/10.48550/arXiv.2304.08485 arXiv:2304.08485 [cs].

[30] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. SSD: Single Shot MultiBox Detector. Vol. 9905. 21–37. https://doi.org/10.1007/978-3-319-46448-0_2 arXiv:1512.02325 [cs].

[31] Mark Nixon and Alberto Aguado. 2019. *Feature Extraction and Image Processing for Computer Vision*. Academic Press.

[32] OpenAI, Josh Achiam, Steven Adler, et al. 2024. GPT-4 Technical Report. https://doi.org/10.48550/arXiv.2303.08774 arXiv:2303.08774 [cs].

[33] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. DINOv2: Learning Robust Visual Features without Supervision. https://doi.org/10.48550/arXiv.2304.07193 arXiv:2304.07193 [cs].

[34] J. R. Parker. 2010. *Algorithms for Image Processing and Computer Vision*. John Wiley & Sons. Google-Books-ID: BK3oXzpxC44C.

[35] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding Multimodal Large Language Models to the World. https://doi.org/10.48550/arXiv.2306.14824 arXiv:2306.14824 [cs].

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. [n. d.]. Learning Transferable Visual Models From Natural Language Supervision. https://doi.org/10.48550/arXiv.2103.00020 arXiv:2103.00020 [cs]

[37] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. http://arxiv.org/abs/2204.06125 arXiv:2204.06125 [cs].

[38] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. https://doi.org/10.48550/arXiv.2102.12092 arXiv:2102.12092 [cs].

[39] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 779–788. https://doi.org/10.1109/CVPR.2016.91 ISSN: 1063-6919.

[40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. https://doi.org/10.48550/arXiv.1506.01497 arXiv:1506.01497 [cs].

[41] Stefan Rüger and Gary Marchionini. 2009. *Multimedia Information Retrieval*. Morgan & Claypool.

[42] Ryo Suzuki, Mar Gonzalez-Franco, Misha Sra, and David Lindlbauer. 2023. XR and AI: AI-Enabled Virtual, Augmented, and Mixed Reality. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*

(UIST '23 Adjunct). Association for Computing Machinery, New York, NY, USA, 1–3. https://doi.org/10.1145/3586182.3617432

[43] Gemini Team, Rohan Anil, Sebastian Borgeaud, et al. 2024. Gemini: A Family of Highly Capable Multimodal Models. http://arxiv.org/abs/2312.11805 arXiv:2312.11805 [cs].

[44] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. https://doi.org/10.48550/arXiv.2302.13971 arXiv:2302.13971 [cs].

[45] Lucia Vadicamo, Rahel Arnold, Werner Bailer, Fabio Carrara, Cathal Gurrin, Nico Hezel, Xinghan Li, Jakub Lokoc, Sebastian Lubos, Zhixin Ma, Nicola Messina, Thao-Nhu Nguyen, Ladislav Peska, Luca Rossetto, Loris Sauter, Klaus Schöffmann, Florian Spiess, Minh-Triet Tran, and Stefanos Vrochidis. 2024. Evaluating Performance and Trends in Interactive Video Retrieval: Insights From the 12th VBS Competition. *IEEE Access* 12 (2024), 79342–79366. https://doi.org/10.1109/ACCESS.2024.3405638 Conference Name: IEEE Access.

[46] Ethan Waisberg, Joshua Ong, Mouayad Masalkhi, Nasif Zaman, Prithul Sarker, Andrew G. Lee, and Alireza Tavakkoli. 2024. Meta smart glasses—large language models and the future for assistive glasses for individuals with vision impairments. *Eye* 38, 6 (April 2024), 1036–1038. https://doi.org/10.1038/s41433-023-02842-z Publisher: Nature Publishing Group.

[47] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. 2024. YOLOv10: Real-Time End-to-End Object Detection. https://doi.org/10.48550/arXiv.2405.14458 arXiv:2405.14458 [cs].

[48] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. NExT-GPT: Any-to-Any Multimodal LLM. https://doi.org/10.48550/arXiv.2309.05519 arXiv:2309.05519 [cs].

[49] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities. https://doi.org/10.48550/arXiv.2308.02490 arXiv:2308.02490 [cs].

[50] Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. 2023. Contextual Object Detection with Multimodal Large Language Models. https://doi.org/10.48550/arXiv.2305.18279 arXiv:2305.18279 [cs].

[51] Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. MM-LLMs: Recent Advances in MultiModal Large Language Models. https://doi.org/10.48550/arXiv.2401.13601 arXiv:2401.13601 [cs].

[52] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7W: Grounded Question Answering in Images. https://doi.org/10.48550/arXiv.1511.03416 arXiv:1511.03416 [cs].

[53] Arzu Çöltekin, Ian Lochhead, Marguerite Madden, Sidonie Christophe, Alexandre Devaux, Christopher Pettit, Oliver Lock, Shashwat Shukla, Lukáš Herman, Zdeněk Stachoň, Petr Kubíček, Dajana Snopková, Sergio Bernardes, and Nicholas Hedley. 2020. Extended Reality in Spatial Sciences: A Review of Research Challenges and Future Directions. *ISPRS International Journal of Geo-Information* 9, 7 (July 2020), 439. https://doi.org/10.3390/ijgi9070439 Number: 7 Publisher: Multidisciplinary Digital Publishing Institute.

A Method for Efficient Structured Data Generation with Large Language Models

Zongzhi Hou
Huawei
Shanghai, China
houzongzhi@huawei.com

Ruohan Zhao
Huawei
Shenzhen, China
zhaoruohan@huawei.com

Zhongyang Li*
Huawei
Shenzhen, China
lizhongyang6@huawei.com

Zheng Wang
Huawei
Singapore, Singapore
wangzheng155@huawei.com

Yizhen Wu
Huawei
Shanghai, China
wuyizhen@huawei.com

Junwei Gou
Huawei
Shanghai, China
goujunwei@huawei.com

Zhifeng Zhu
Huawei
Nanjing, China
zhuzhifeng3@huawei.com

ABSTRACT

With the rapid advancement of large language model technology, the data utilized for training these models has become increasingly significant. The quality of text data samples produced by large unsupervised models is often inadequate, leading to insufficient outcomes. This inadequacy arises from the model's constrained capacity to precisely emulate the underlying structure of the data without direct supervision, resulting in outputs that may lack the necessary fidelity and relevance to the authentic data distribution. In order to overcome the shortcomings of training data generation for specific language generation tasks, this paper proposes a fast data generation system (Fast Data Generation System, FDGS) that can handle multi-modal and structured data generation. As a method for generating data, FDGS uses clustering abstraction to handle multiple data input types through templates. This approach allows for quick data generation and reduces consumption. FDGS is robust, ensuring stable and reliable performance under various conditions. It is more cost-effective in terms of token usage compared to traditional methods that work on a per-instance basis and do not use templates. By abstracting and clustering different input types, FDGS can efficiently generate data from large models. This system is highly adaptable, making it a great choice for multi-modal data generation tasks. It relies on the basic functions of general large-scale language models and employs a query-answer bidirectional generation mechanism to achieve fast data amplification.

CCS CONCEPTS

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

KEYWORDS

Multi-modality, Data Generation, Artificial Intelligence, Large Language Model

ACM Reference Format:
Zongzhi Hou, Ruohan Zhao, Zhongyang Li, Zheng Wang, Yizhen Wu, Junwei Gou, and Zhifeng Zhu. 2024. A Method for Efficient Structured Data Generation with Large Language Models. 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, 9 pages. https://doi.org/10.1145/3688866.3689127

1 INTRODUCTION

The rapid increase in data, now reaching 2.5 exabytes daily, highlights the need for better methods in data management and use [2]. Machine learning and deep learning have used these large datasets to enhance analytical precision and predictive accuracy [30]. However, relying heavily on big data also reveals the flaws in traditional data collection methods, which are often labor-intensive and biased [28]. This problem is even more evident when generating multi-modal data, such as images or LiDAR point cloud data for autonomous driving [23, 24], which usually require a lot of effort to gather through direct collection or simulation tools [15, 53].

In model training and testing, semi-automatic generation of grammar test items using Natural Language Processing (NLP) techniques is common [35]. Additionally, acquiring image or text data is no longer limited to direct retrieval from live networks. Advanced techniques like attribute manipulation based on Large Language Models (LLMs) are now key for creating efficient training datasets. Moreover, synthetic data generation platforms like DataDreamer are essential, enabling the creation of reproducible and customized datasets for specific research or operational needs [39]. Understanding the different methods for data generation—including real-world data collection, synthetic data creation, and using large models—is

[*] Corresponding author

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

ACM ISBN 979-8-4007-1193-0/24/10
https://doi.org/10.1145/3688866.3689127

crucial for informed decision-making and advancing computational sciences and modeling. Key considerations include data distribution nuances, domain scope, and alignment with practical end-user needs, which underscore the importance of choosing effective data generation strategies for various applications [41].

The first method for data generation involves collecting real data from online networks, which provides the advantage of capturing observational data in its natural form [16]. Real data from various online platforms like social networks, websites, and databases offers raw insights into complex phenomena. This method is vital for studies needing high ecological validity, especially in social science research involving human behavior, where accurate representation of real-world contexts is essential [29]. However, this approach has notable downsides due to potential inaccuracies and biases in the data sources. Online data can be skewed by over-representation of certain demographic groups [36], and privacy concerns require careful ethical considerations when collecting personal data [56].

The second method focuses on generating data through large models, which replicate defined data characteristics like distribution, domain range, and alignment with actual user queries [19, 48]. Nevertheless, a major challenge with this method is its dependence on the model's accuracy and inclusiveness. Discrepancies within the model can result in ineffective or biased data generation [4], and developing these models requires extensive expertise in modeling, which might be restrictive for some organizations [5].

Synthetic data generation utilizes algorithms and statistical methods to create artificial data that mimics real-world patterns [12]. This technique's primary advantage is its ability to avoid privacy issues associated with real data since the generated data is entirely artificial and not subject to personal data regulations [1]. Additionally, synthetic data can be customized to suit specific research queries, enhancing representation of certain scenarios or populations [52]. However, the limitations of synthetic methods often include the inability to mass-produce high-quality data for model training [18].

To address these challenges, this paper proposes a Data Generation System, an innovative solution for multi-modal data generation strategies. The Fast Data Generation System (FDGS) effectively handles various types of data inputs and structured data by employing a clustering abstraction method supporting diverse data input types, thus addressing the deficiencies of large unsupervised models. Data generation begins with images, LiDAR point clouds, and text input as three multi-modal data sources. Through preliminary image recognition and point cloud clustering, raw data is abstracted into patterns [15, 49]. Based on these patterns, large models are used for automated batch generation, and with self-supervision, the system continuously optimizes and adjusts prompts to ensure data generation aligns with preset patterns [9].

Recent advancements in machine learning for synthetic data generation have shown potential in mitigating data scarcity and privacy concerns [38]. Traditional methods, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), have been crucial in generating realistic data across various domains [18, 26]. However, these methods often struggle with the complexity and diversity of multi-modal data inputs [21].

This paper's contributions are threefold: The contributions of this paper are threefold:

(1) **Innovative Data Generation Approach:** FDGS introduces a novel data generation methodology that leverages the strengths of multi-modal data handling and structured data generation, providing more accurate and inclusive datasets for model training [34].
(2) **Efficiency and Cost-Effectiveness:** FDGS reduces token usage and uses a bidirectional generation method, providing a cost-effective solution that lowers the expenses of data collection and storage [7].
(3) **Reducing Bias and Improving Generalization:** FDGS is robust and adaptable, which helps reduce biases and improve model generalization, ensuring the generated data is varied and reflects real-world scenarios [4].

Our proposed Fast Data Generation System (FDGS) represents a new approach in this field. FDGS excels in handling various types of data, such as images, LiDAR point clouds, and text inputs. Using clustering abstraction methods, FDGS effectively processes different input types and patterns, creating structured data that matches predefined patterns [15]. This approach addresses the inefficiencies often seen in large unsupervised models when dealing with diverse data formats.

The combination of clustering abstraction and large model-based pattern generation allows FDGS to automate the data generation process. Self-supervised learning techniques help the system continuously optimize prompts, ensuring that the generated data consistently meets the desired patterns [9]. This method not only improves the efficiency of data generation but also enhances the quality and relevance of the synthetic data produced.

In summary, FDGS provides a practical solution for multimodal data generation, addressing key gaps in current methods. Its application has the potential to transform various fields, including autonomous driving, medical imaging, and natural language processing, where the diversity and complexity of data inputs are crucial [30].

2 RELATED WORK

This section reviews research efforts related to EDGS, focusing on automatic prompt optimization and LLM-based intelligent data generation.

2.1 Automatic Optimization of Prompts

Prompt optimization is crucial for large language models (LLMs) to generate accurate results [45]. This paper presents a new approach that imitates gradient-based optimizers. AutoPrompt refines prompts using model gradients, enhancing LLMs' performance in knowledge retrieval tasks [50], outperforming manually crafted prompts [45]. However, this method primarily applies to text models, with limited effectiveness for multimodal inputs like images and point cloud data.

The Gradient-inspired LLM-based Prompt Optimizer (GPO) improves on traditional prompt optimization by learning from past data, combining generation-based refinement with cosine-based control. This approach enhances precision and relevance, achieving up to 56.8% improvement on the Big-Bench Hard dataset and 55.3% on MMLU [13].

AutoHint enhances prompts by providing task-specific hints, using clustering and balanced sampling strategies. This method improves accuracy across multiple tasks, particularly in zero-shot settings, by applying structured hints to improve LLM prompts [32].

Implicit Reflection-based Prompt Optimization refines prompts using historical data, enhancing performance without direct reflection. By analyzing past prompts and scores, it optimizes LLM tasks effectively [44]. While these approaches improve LLM performance in text domains, they lack effectiveness in optimizing prompts for multimodal inputs like images and point clouds.

2.2 Bidirectional Data Generation System

Automated data generation is vital across various fields, enabling advancements in machine learning, computer vision, NLP, and software engineering. This section reviews significant contributions and methodologies related to automated data generation.

In machine learning, automated data generation is crucial when labeled data is limited. Techniques like rotation, translation, and GANs enhance model robustness. Goodfellow et al. (2024) introduced GANs, where a generator creates synthetic data and a discriminator distinguishes it from real data, improving robustness despite challenges like mode collapse [17, 54, 55].

Radford et al. (2024) expanded on GANs with Deep Convolutional GANs (DCGANs) for image synthesis, improving image quality and diversity [40]. However, DCGANs may produce images with artifacts.

Conditional GANs (cGANs) introduced by Isola et al. (2024) advanced image-to-image translation tasks, useful in generating images or semantic segmentations, though limitations remain for high-resolution applications [20]. Variational Autoencoders (VAEs) generate diverse images from latent distributions but often produce blurry results [27].

In NLP, models like BERT and GPT-3 revolutionized text generation. BERT uses bidirectional transformers for contextually appropriate text but requires substantial computational resources [10]. GPT-3 generates high-quality text across contexts with minimal examples but faces challenges in controlling output quality and addressing biases [6].

In software engineering, techniques like symbolic execution and fuzzing are used for test case generation and vulnerability identification. Cadar and Engler (2024) discussed symbolic execution for comprehensive program behavior coverage, though scalability remains a challenge [8]. Sutton et al. (2024) explored fuzzing, a technique for generating test cases to discover software bugs and security vulnerabilities. They outlined various fuzzing strategies, but challenges include generating meaningful test cases for complex software and reducing false positives, impacting the overall efficacy [47].

Despite progress, generated data still faces quality and diversity issues. Liu and Yu (2024) proposed methods for improving synthetic data fidelity under constraints [33]. Most recent works struggle with multimodal input handling and generating data in specific directions. The proposed EDGS overcomes these limitations by preprocessing input content as patterns and rapidly generating text based on these patterns.

Ethical considerations are also critical in data generation. Barocas and Selbst (2024) examined ethical implications and biases, proposing frameworks to address these challenges while recognizing the difficulty of eliminating biases entirely [3].

3 EFFICIENT DATA GENERATION SYSTEM

In this section, we discuss the methodology in Efficient Data Generation System, including data construction, training preparation and evaluation benchmark.

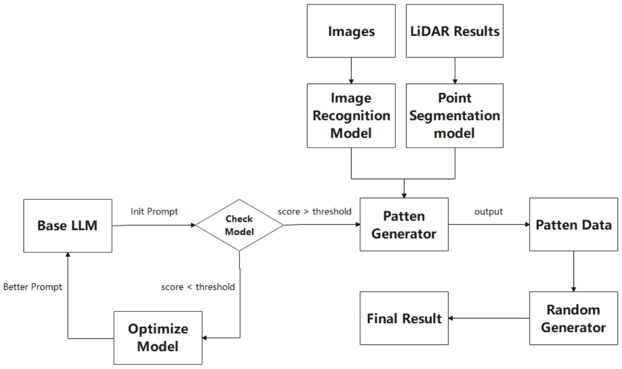


Figure 1: Data Generate System

3.1 Data Construction

The construction of a robust data-set is a critical foundational step in the training of large machine learning models. The quality, volume, range, diversity, and relevance of the data directly affect the model's performance. This section discusses in detail the methodology for generating and preparing training data suitable for large-scale models focusing on deep learning.

- Data Collection
  Carla simulation software is a professional autonomous driving simulation software developed based on the Unreal engine. Epic Games constructed the Unreal engine as a game engine, integrating rendering, collision detection, AI, graphics, and networking under the engine [11]. The Unreal engine has excellent performance in simulating the driving environment of the vehicle and possible collisions during driving. Therefore, Carla simulation software can approximate the real scene regardless of the scene's reality and the sensor's simulation.
  Although Carla has the above advantages, it is in continuous development and improvement, so procedural bugs sometimes appear, which may cause the running script to crash. In addition, the ROS interface supported by Carla is not highly supported on the Windows platform, so some extensions need to be converted to the Ubuntu 18.04 system for improvement.
  In this project, I use the 0.9.11 version of the Carla simulator [1]. This software is scripted based on Python to realise the

[1] We utilized the Carla open-source tool (version 0.9.11) to generate experimental data. This tool is licensed under the MIT License, which allows us to freely use, modify, and distribute the source code. Please visit https://carla.org/services/ to access the full text of the license.

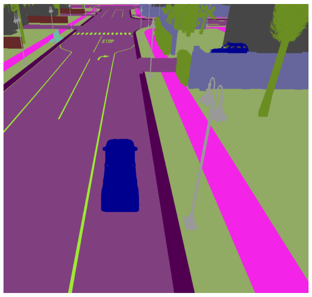


Figure 2: Images by a semantic segmentation camera

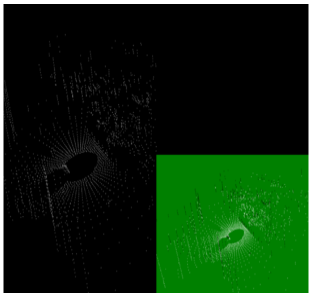


Figure 3: LiDAR information

required functions. It also supports the Windows platform and Linux system. In terms of software, the Carla simulator only supports Python. Considering the compatibility of OpenCV and other aspects, install Python3.7 as the software environment for scripting.

First, the multi-modal data content is collected from maps and driving facilities built by the autonomous driving simulation software Carla. Through the installation of laser radar and RGB camera, semantic segmentation camera on the virtual simulation vehicle, the image information and laser radar information around the vehicle are collected, and the diversified information is obtained for experimental verification. The example is shown in the figure.

The process begins with an extensive data collection phase. Data can be sourced from various repositories, open-source datasets, or through web scraping, APIs, and sensor network deployments, considering ethical and legal constraints [16]. Heterogeneous data types are collected to enrich the dataset, including structured data (tables, CSV files) and unstructured data (text documents, social media posts) [35].

- Annotation and Labeling Once the data is collected, it must be properly annotated or tagged to provide a true reference for the model. We classify scenarios based on actual driving scenarios and mark information such as vehicles, traffic lights, and pedestrians. The labeled data is used for visual model training to combine the point cloud clustering result and visual recognition result to generate the target pattern [15].
- Data Cleaning Data cleaning is essential to remove noise and correct errors in the dataset. It involves tasks like deduplication, dealing with missing values, and removing outliers or irrelevant instances. The laser radar data is easy to be interfered with, so Gaussian filtering method is used to remove the laser point cloud noise [15]. The image is obtained by filtering image noise, converting the image into a grayscale image, and then segmenting the image to reduce the impact of noise. For text input, data cleaning mainly includes filtering irrelevant content in the actual corpus through a vocabulary and removing excessively long or excessive phrases from the input corpus [35].
- Feature Engineering Feature engineering transforms raw data into a format that machine learning algorithms can work with more effectively. It may involve data transformations such as normalization, scaling, tokenization for text, or feature extraction techniques to reduce dimensionality [30]. Encoding of categorical variables, one-hot encoding for high cardinality features, or embeddings for deep learning models are considered.
- Data Splitting The assembled dataset is divided into training, validation, and testing sets. A common split ratio is 70% for training, 15% for validation, and 15% for testing, although this can vary depending on the dataset size and the model requirements. Stratified sampling may be used to preserve the distribution of classes.
- Data Augmentation Data can be sourced from various repositories, open-source datasets, or through web scraping, APIs, and sensor network deployments, considering ethical and legal constraints. Heterogeneous data types are collected to enrich the dataset, including structured data (tables, CSV files), untextual data (images, videos), and unstructured data (text documents, social media posts) [46].

3.2 Training preparation

Our experiment preparation includes model preparation, data generation, and evaluation of the final generation result. Because the Prompt of different models has a self-tuning link, we assume that effective data output can be obtained after the automatic adjustment of the Prompt is completed [45]. Next, we will compare the data generation speed and token consumption of different data construction methods.

3.3 Evaluation benchmarks

In the domain of computational models, particularly those involving large-scale architectures, benchmarks play a fundamental role in assessing the effectiveness, efficiency, and overall performance of algorithms, systems, or methodologies under study. The utilization of benchmarks follows rigorous scientific methods to ensure that the data obtained is valid and reliable. In this context, we partition the dataset into training, validation, and test sets to ensure that multiple pattern extraction models and language generation systems not only undergo thorough training but also receive validation to prevent overfitting. Ultimately, performance is evaluated through rigorous testing to gauge the system's overall effectiveness [30].

Benchmarks typically encompass a variety of metrics related to the final objectives of the model. In this study, visual metrics are employed to assess the accuracy of visual target evaluations, while point cloud segmentation metrics focus on the proportion of clustered points within the original object point clouds. For text corpus information, the ROUGE model is utilized to evaluate the accuracy, precision, and recall of structured information [31]. Beyond mere predictive performance, benchmarks also assess computational aspects such as inference time, training efficiency, memory usage, and scalability [30]. Given the complexity of our system, which involves multiple modules, a full-scale evaluation of data processing stages is not conducted. Instead, we focus on evaluating the core generation system, considering robustness, interpretability, fairness, and ethical implications to provide a comprehensive assessment [14].

The benchmarking process is methodical and structured to ensure detailed analysis and unbiased conclusions. Based on our research hypothesis—that the EDGS (Feature-based Data Generation System) can enhance the efficiency of generating multi-modal input data—we integrate model design with benchmark standards. In the training phase, we develop and train models on the training set, utilizing the YOLO-v5 model for visual tasks due to its rapid inference capabilities [22]. For laser point clouds, clustering models are trained to perform spatial clustering calculations, extracting pattern information from point clouds. This information is then integrated with image data to unify point cloud and image recognition results with structured textual descriptions.

By employing these benchmarks, we guarantee a complete and impartial evaluation of the system's performance and potential improvements. This way of doing things helps quantify how well the system works, along with other factors like computational efficiency or model robustness and even ethical considerations. Our visual model can reach 97% accuracy for item recognition.

3.4 Training Pattern Model

In the domain of machine vision algorithms, the choice is very important for the efficiency and accuracy of autonomous driving systems. The KITTI dataset itself is a very good benchmark for evaluating computer vision techniques in autonomous driving, providing a complete platform to test how well different algorithms perform [15]. It was created by the Karlsruhe Institute of Technology and the Toyota Technological Institute at the University of Stuttgart and comprises a large number of vision tasks: stereo, optical flow, visual odometry, and object detection and tracking.

In our journey through the world of machine vision algorithms, we came across the YOLO (You Only Look Once) series [2], which though not the fastest or most accurate, strikes a good balance between both. The YOLO series, specifically Yolo-v5, achieves an excellent compromise between on-the-fly processing ability and high detection accuracy, important in meeting strict latency needs in autonomous systems [42]. The training data for our vision recognition model was collected with utmost care and then annotated by going through RGB camera feeds and semantic segmentation camera outputs [15].

In the field of LiDAR point cloud processing, extracting individual objects or regions from laser-scanned point cloud data is not a trivial task. We used the Euclidean Cluster Extraction algorithm. It is very good at forming clusters by connecting adjacent points, unlike K-Means clustering or Region Growing [43]. With a simple adjustment of minimum cluster size, this algorithm is able to segment vehicle and pedestrian point clouds very well, while filtering ground, walls, and trees out as noise. What makes it unique is its ability to very quickly process uniform point cloud data; that is why it is so good for our system.

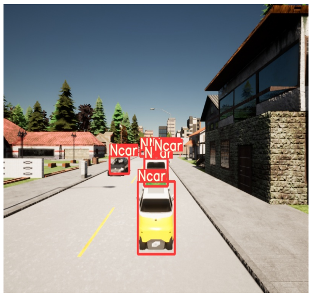


Figure 4: Yolo-v5 image recognition result

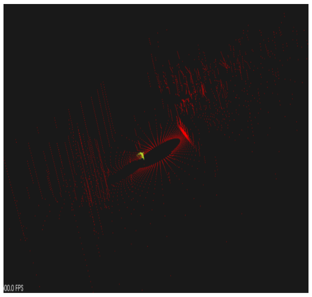


Figure 5: Point cloud for a car clustering result

For the language model component, we constructed a language training dataset by processing queries and utilized the Pangu-13B model as our foundational model, subjecting it to Supervised Fine-Tuning (SFT) [51]. Throughout the training regimen, we adhered to the principle of incremental learning, which replaced the need for parameter updates across the entire dataset. This approach achieved a harmonious balance between iteration speed and convergence stability. To mitigate overfitting, our dataset incorporated regularization techniques, dropout, and batch normalization, complemented by early stopping based on performance metrics derived from a validation set. Hyperparameter tuning was conducted using the Adam optimization method, which autonomously identified the optimal learning rate [25]. The training process involved the propagation of network batches, gradient computation, and weight adjustments, reiterated over multiple epochs until a stable generation level was attained.

Through these methodologies, our model training yielded the following results in image recognition and LiDAR clustering. By harnessing the comprehension capabilities of large models and integrating textual information, we effectively transformed multi-modal inputs into structured textual information that is more digestible for the model. This includes object types, positions, and descriptions, ultimately generating structured information in textual form.

4 EXPERIMENTS

In this part, we will introduce the following research questions by a series of experiments:

Q1 Measure the number of rounds consumed by different methods to generate the same number of results with the same initial Prompt input and assuming that the optimal threshold is reached.

Q2 Measure the number of tokens consumed by different methods to generate the same number of results with the same initial Prompt input and assuming that the optimal threshold is reached.

[2] In this study, we utilized Yolo-V5, an open-source tool released under the AGPL-3.0 license, to generate our experimental data. This license permits users to freely use, modify, and distribute the source code, provided that they adhere to the terms and conditions of the license. In accordance with the requirements of AGPL-3.0, if the outcomes of this study are employed as part of a network service, we provide access to the source code of the tool to ensure transparency and uphold the spirit of open-source. The full text of the AGPL-3.0 license can be accessed at https://github.com/ultralytics/yolov5. Please note that no warranties of any kind are provided with the use or distribution of Yolo-V5, and we assume no legal liability arising from its use.

4.1 Experimental Setup

In this section of our paper, we provide comprehensive details of the methods and tools used during our research.

First, we start by outlining our main goal, which is to reduce token consumption while increasing production speed by building a system that produces targeted high-quality data in bulk. Specifically, we aim to assess the effectiveness of large models in generating high-quality data quickly. Key aspects include comparison of different models, variation of parameters, and accuracy of generated results as a system standard measure.

Next, we introduce the big data model used in the experiment. These large models are selected for classification based on their differentiated parameters and features. Notably, we use models that have been recently developed or updated in the data science field, including models such as GPT-3.5, GPT-4, and Mistrol-7B [7]. By including a variety of models, we can get a broader perspective on how each model performs uniquely. Considering the requirements of each model on the prompt, the fixed instruction + task description is used as the prompt.

In addition, the data sources and parameters we apply to these models play an important role. Since our paper is based on data generation, we focus on synthetic datasets that are close to real-world data, explaining why these datasets were chosen, their size, and what they represent [52].

In terms of software and hardware, the experiment is based on the platform cloud service system, equipped with advanced GPUs and capable of handling large-scale model requirements [51].

Validation procedures are another key factor. We list the measures used to assess the quality of the data generated, such as precision, recall, F1-Score, ROC curve, and other techniques. Benchmark models or standards for measuring the performance of large models are also discussed [31].

The experimental setup section explains how the whole experimental process was taken to ensure a standardized and unbiased approach, and also the time required to evaluate the feasibility of using such large-scale models in the real world [30].

In conclusion, this section provides a clear understanding of the methodology concerning our research, which not only gives a basis for our findings but also serves as a reference for other researchers in the field.

4.2 Model Rounds Evaluation

In our experiments, we evaluate different data generation methods and compare them with large-scale models. Depending on the starting point, other techniques may take more iterations to achieve the desired data. We explore and explain this effect, providing both quantitative and qualitative insights.

The data generation process started with irregular queries and regular results as our main goal. The query formulated was "Please help me remember xxx." This query initiates interaction with the large-scale model. The result generated is represented in JSON format, with content such as "subject" and "address" as key-value entries. The output is: "subject":"xxx,""address":"xxx..."

We set three benchmarks for evaluating the efficiency of our data system: generating one thousand, five thousand, and ten thousand pieces of data. These benchmarks allow us to test, understand, and evaluate the effectiveness and efficiency of our systematic approach to data generation. Each benchmark provides unique insights into large-scale data generation.

The basic hypothesis of the experiment was to "measure the number of rounds consumed by different methods to generate the same number of results using the same initial cue input and assuming that the optimal threshold is reached." We examine whether different data generation methods have a substantial impact on the efficiency of generating structured results within set prompt and iteration limits.

The number of rounds consumed by different data generation methods is the key metric. The goal is to compare the efficiency and effectiveness of each method, assuming that the optimal threshold is reached.

We first keep the initial prompt input the same in all methods. Consistent inputs allow us to remove any anomalies or discrepancies, ensuring a clear comparison [45].

The optimal threshold assumption is crucial for our experiment. We assume that each method reaches a point of optimal efficiency after a certain number of rounds. Assuming all methods reach this threshold, we focus on the efficiency of the methods.

We use the number of rounds to achieve the final goal as a characteristic indicator to measure system generation efficiency. The experimental results are shown in Table 1.

We measure four distinct approaches: Bulk Generation, Maximal Generation by the Model, EDGS (No Mixed), and EDGS (Mixed), towards achieving the target quantity of accurate entries generated by the GPT-4 model [37].

Bulk Generation involves providing the input prompt to the large-scale model, obtaining one result per round. This method focuses on generating one accurate result at a time [7].

In contrast, Maximal Generation by the Model attempts to gather as many accurate results as possible per round, with an upper limit of approximately 50 entries per round. We measure the efficacy based on the average number of rounds taken to meet the desired entry limit.

EDGS (No Mixed) implements the Efficient Data Generation Strategy (EDGS) to construct templates and corresponding slot information, generating approximately 50 place data entries per cycle.

EDGS (Mixed) builds upon Method C with an incorporated mix of places and templates. The generation process halts once the accumulated templates ensure avoidance of duplicate data.

This study gauges the efficiency of these methods in generating precise and structured data using the GPT-4 model, contributing valuable insights for large-scale data generation [37].

Table 1: Model Rounds for Different Data Generate System

Method	1000-items	2000-items	5000-items	10000-items
Bulk Generation	1120	2240	5600	11200
Maximal Generation by the Model	224	448	1120	2240
EDGS (No Mixed)	224	448	1120	2240
EDGS (Mixed)	224	448	560	560

Due to factors such as model hallucinations and duplicate outcomes, data generated by models often can't be directly utilized; we have incorporated these scenarios in our failure conditions for comprehensive consideration [4]. Drawing on information from both practical experiments and descriptions within other studies, under normal conditions with automated optimization of Prompt guidance for large model result generation, approximately 12% of the results get discarded due to reasons such as repetition or noncompliance with data generation norms [7]. In our results, we will also consider this part in terms of the number of rounds and actual consumption. Consequently, we will accommodate these failure scenarios when evaluating the data-generating capabilities of different models.

4.3 Tokens Consumption Evaluation

Continuing from the previous scenario, this portion of our study entails a comprehensive assessment of the number of tokens consumed by the same methods (Bulk Generation, Maximal Generation by the Model, EDGS (No Mixed), and EDGS (Mixed)) to generate an equal number of results with the same initial Prompt input. This evaluation, unlike the prior one, switches the metric from the number of rounds to the number of tokens utilized, once the optimal threshold is supposed to have been attained, thereby offering a different metric for comparison [7].

In Bulk Generation, which achieves results through sequential generation on a large-scale model via an iterative querying process, one result feedback is obtained per round. The token consumption here is relatively simple and linear, with each round potentially consuming a constant number of tokens based on the complexity of the prompt and response.

Maximal Generation by the Model, in comparison, tries to extract multiple accurate outputs per round to the defined upper limit of approximately 50 items in order to maximize efficiency, while still adhering to the principle of accuracy. Although this approach might increase the token consumption per round due to the larger volume of data being generated, it could potentially reduce the overall number of tokens needed to reach the target output quantity, owing to fewer iterations.

EDGS with no mixed data adopts a structured data creation approach—Efficient Data Generation System. It can rapidly generate pattern data and a case database, which might save some tokens as it reuses templates and place or other typical data to create the results, thus reducing the total tokens consumed to reach the desired amount of results.

Unlike EDGS with no mixed data, EDGS with mixed data introduces an element of diversification in the data, combining places and templates in order to avoid data redundancy, which could lead to a slight increase in token consumption. However, this method ensures the novelty and diversity of the generated dataset.

Table 2: Token Consumes for Different Data Generate System.

Method	1000-items	2000-items	5000-items	10000-items
Bulk Generation	112000	224000	560000	1120000
Maximal Generation by the Model	22400	44800	112000	224000
EDGS (No Mixed)	22400	44800	112000	224000
EDGS (Mixed)	22400	44800	56000	56000

This token-based performance evaluation provides a meaningful standpoint to understand and compare the efficiency of the different generating methods in terms of their throughput within a given token limit, presenting another perspective for the practical applicability of these techniques.

4.4 Discussion

Large-scale data models have become increasingly essential in predictive analytics and other data-driven fields of research [7]. Our exploration into the efficacy and efficiency of these models has led us to several consequential findings, echoing and expanding upon the existing body of work that suggests that while large models can be powerful, they are not without their challenges. The comprehensive evaluation of various data generation methods utilizing the same large-scale model has yielded fascinating insights. The four distinct strategies of Bulk Generation, Maximal Generation by the Model, EDGS (No Mixed), and EDGS (Mixed), have shown considerable variability in their efficiency and effectiveness [7, 37].

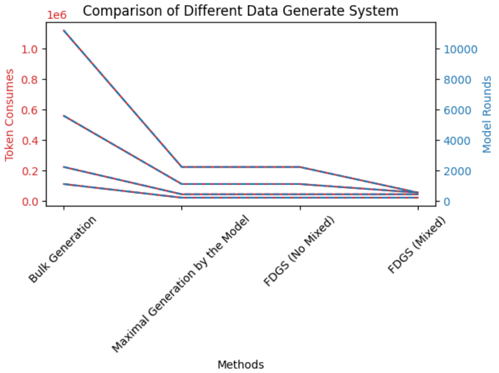


Figure 6: Comparison of Different Data Generate System

However, one major drawback of this approach is that it is relatively inefficient when large volumes of data are needed quickly. The Bulk Generation speed is slower compared to other methods, as it consumes many tokens for each data generation round. This predictability of performance is what is emphasized by Brown et al (2020) [7].

However, one major drawback of this approach is that it is relatively inefficient when large volumes of data are needed quickly. The Bulk Generation speed is slower compared to other methods, as it consumes many tokens for each data generation round. This predictability of performance is what is emphasized by Brown et al (2020) [7].

In order to address some of the limitations of Bulk Generation, the Maximal Generation by the Model method was explored. This approach has proven to be a more viable alternative that allows larger quantities of data to be generated per round while still maintaining high quality in the data output [7]. By maximizing the data output in each round and ensuring its quality, the model significantly brings down the number of required rounds and tokens.

The Maximal Generation by the Model method is a powerful tool for token-efficient large-scale data production since it can generate huge amounts of data without consuming many tokens. This makes it ideal for situations requiring immediate large volumes of data, as token budget limitations will not impede its efficiency. This efficiency is markedly superior, as demonstrated by Brown et al. (2020) [7].

However, this efficiency does come at a cost. The production of data on bigger scales does bring up unique challenges and considerations. Although maximal generation by the model is efficient, it has to be well controlled to work on these issues effectively. One of the key among them would be how to handle the resultant dataset and ensure that the data produced is authentic and of good quality.

EDGS (No Mixed) employs data templates and an efficient data generation strategy to expedite data production. The reusing of templates and place data entries significantly reduces token usage and the number of rounds required, except the model needs to ensure the avoidance of duplicate entries and the occasional addition of new templates [7].

EDGS (Mixed) further refines the technique by introducing a blend of diverse templates and places in the data generation process. This ensures a unique and varied dataset. Undeniably, this method necessitates extra carefulness to avoid data redundancy.

It is evident from the above results that all four data generation methods have their unique advantages—Bulk Generation serves predictability, Maximal Generation by the Model prioritizes quantity while ensuring quality, EDGS (No Mixed) brings the speed, and EDGS (Mixed) guarantees speed as the most fast method [7].

Furthermore, these methods' efficiencies greatly depend on the initial input prompt. BThe bias in the initial prompt might tilt the results, indicating how important it is for the rest of the process. Consequently, it is very necessary to take into account that the initial prompt should be unbiased and not be inclined towards any specific method [45].

When evaluating these techniques' efficiencies, one must also consider the optimal threshold assumption, which states that each method reaches a point of optimal efficiency after a certain number of iterations. Notably, the efficiency of methods may vary before and after hitting this point, underscoring the importance of this consideration.

This detailed analysis explains the process of how to generate a large amount of data by using large-scale models. It discusses what should be taken into account and what difficulties researchers may have while trying to make their data generation procedures effective, correct, and prompt.

The results of these experiments contribute not only to new knowledge about various methods of data generation but also indicate that it is possible to delve further into this process and improve it. Subsequent studies should perfect the methods identified here, as well as try to surpass the frontiers of data generation with large-scale models. There is little explored possibility yet in which hybrid models can be tested for the good points of all four techniques; this can create some unprecedented efficiencies in data generation.

5 CONCLUSION

In conclusion, this in-depth study of the Efficient Data Generation System (EDGS) serves as valuable input for understanding and developing more efficient and effective user-centric data generation methodologies. It not only unveils the potential that current methodologies have but also opens up new avenues of exploration in data generation optimization, especially in the era where data plays such a fundamental role [2].

EDGS thrives on extracting structured information and utilizing template sentences for bulk results production. Hence, it performs well when there's a need for abstract data. Abstract data refers to key slot values that can be utilized for abstraction. On the flip side, for non-abstract results that are not amenable to template abstraction—like unordered information or a non-replaceable code structure—the abstract data bulk generation doesn't present a distinct advantage [7].

Furthermore, since bulk-generated outcomes include content replaced in bulk, when it is actually used for model training, its effectiveness might be slightly less successful compared to data produced manually or data generated individually in the same quantity. This is because bulk replacement can lead to some level of repetition.

In our experimental phase, Efficient Data Generation System (EDGS) integrates structured data, completes data generation in the fewest rounds, and consumes fewer tokens [7]. However, like all methods, this one has its limitations. EDGS: It lacks support for scenarios with high continuity requirements such as voice and video, and multi-modal input depends on the effect of front-end image and point cloud clustering [15, 49]. This is because the current multi-modal large model has limitations in understanding patterns of different scenario-based inputs, and cannot directly abstract key patterns in different forms of inputs and replace and quickly generate the key patterns. In future work, we will try to train information recognition models (such as image models, point cloud clustering models, etc.) into the large model, so that the large model has the ability to directly understand multi-type inputs.

Despite these drawbacks, in specific circumstances that demand rapid acquisition of a vast amount of structured data, EDGS can virtually contribute towards swift bulk output generation. Additionally, it can lessen the consumption of model tokens.

Lastly, The ability of EDGS to reduce model token consumption highlights an important aspect of data generation: optimizing resources. In the long-term future, we will continue to study the research on the rapid data construction of large multi-modal input models, and how to efficiently build patterns to quickly generate relevant data.

REFERENCES

[1] John M. Abowd. 2018. The U.S. Census Bureau Adopts Differential Privacy. *Proceedings of the National Academy of Sciences* 116, 20 (2018), 9848–9851.

[2] Ruchita H Bajaj and P Ramteke. 2014. Big data–the new era of data. *International Journal of Computer Science and Information Technologies* 5, 2 (2014), 1875–1885.

[3] Solon Barocas and Andrew D Selbst. 2024. Big Data's Disparate Impact: Addressing the Ethical Implications. *Ethics and Information Technology* 26, 1 (2024), 45–59.

[4] Emily M. Bender et al. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021), 610–623.

[5] Rishi Bommasani et al. 2021. Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258* (2021).

[6] Tom Brown et al. 2024. Language Models are Few-Shot Learners: Scaling Up GPT-3. *Journal of Artificial Intelligence Research* 70 (2024), 245–261.

[7] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[8] Cristian Cadar and Dawson R Engler. 2024. KLEE: Unassisted and Automatic Generation of High-Coverage Tests for Complex Systems Programs. In *Proceedings of the 15th International Conference on Architectural Support for Programming Languages and Operating Systems.* 209–220.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. *International conference on machine learning* (2020), 1597–1607.

[10] Jacob Devlin et al. 2024. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Transactions of the Association for Computational Linguistics* 12 (2024), 345–361.

[11] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. Carla: An open urban driving simulator. *Conference on Robot Learning* (2017), 1–16.

[12] Awad El et al. 2021. Machine Learning for Synthetic Data Generation: A Review. *Journal of Machine Learning Research* 22, 1 (2021), 1–35.

[13] Tianyu Gao et al. 2020. Making Pre-trained Language Models Better Few-shot Learners. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 3816–3830.

[14] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2013. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research* 32, 11 (2013), 1231–1237.

[16] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457, 7232 (2009), 1012–1014.

[17] Ian Goodfellow et al. 2024. Generative Adversarial Networks: Recent Advances. *Journal of Machine Learning Research* 25, 1 (2024), 1–25.

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[19] Donald Hernandez et al. 2021. Measuring and Increasing the Usefulness of Interpretable Models. *Journal of Artificial Intelligence Research* 72 (2021), 233–248.

[20] Phillip Isola et al. 2024. Image-to-Image Translation with Conditional Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1125–1134.

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), 1125–1134.

[22] Glenn Jocher, Akash Chaurasia, and Jirka Stoken. 2021. YOLOv5 by Ultralytics. *GitHub repository* 6 (2021).

[23] Ce Ju, Zheng Wang, Cheng Long, Xiaoyu Zhang, and Dong Eui Chang. 2020. Interaction-aware kalman neural networks for trajectory prediction. In *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 1793–1800.

[24] Ce Ju, Zheng Wang, and Xiaoyu Zhang. 2018. Socially aware kalman neural networks for trajectory prediction. *arXiv preprint arXiv:1809.05408* (2018).

[25] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[26] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[27] Diederik P Kingma and Max Welling. 2024. Auto-Encoding Variational Bayes: Advances and Applications. *Neural Computation* 36, 2 (2024), 289–305.

[28] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The parable of Google Flu: traps in big data analysis. *Science* 343, 6176 (2014), 1203–1205.

[29] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. 2009. Computational social science. *Science* 323, 5915 (2009), 721–723.

[30] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[31] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. *Text summarization branches out: Proceedings of the ACL-04 workshop* 8 (2004), 74–81.

[32] Paul Pu Liu, Weizhe Yuan, Jun Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. What makes good in-context examples for GPT-3? *arXiv preprint arXiv:2101.06804* (2021).

[33] Xin Liu and Yang Yu. 2024. Recent Advances in Synthetic Data Generation for Machine Learning. *Artificial Intelligence Review* 58, 3 (2024), 367–389.

[34] John Mitchell et al. 2021. Gecko: Bridging Traditional and Deep Learning Models for Tabular Data. *Journal of Data Science* 29, 4 (2021), 567–590.

[35] Ruslan Mitkov, Le An Ha, and Nikiforos Karamanis. 2003. Computer-aided generation of multiple-choice tests. *Proceedings of HLT-NAACL 2003 workshop on Building educational applications using natural language processing* (2003), 17–22.

[36] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. 2013. Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)* 400 (2013), 2013.

[37] OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).

[38] Donggeun Park, Donggyu Hwang, and Changick Kim. 2019. Data augmentation for object detection via progressive and selective instance-switching. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), 3039–3048.

[39] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. 2016. The synthetic data vault. *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (2016), 399–410.

[40] Alec Radford et al. 2024. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 4 (2024), 869–880.

[41] Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Re. 2017. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment* 11, 3 (2017), 269–282.

[42] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), 779–788.

[43] Radu Bogdan Rusu and Steve Cousins. 2010. PCL: Point cloud library. *IEEE International Conference on Robotics and Automation* 9, 1 (2010), 1–6.

[44] Timo Schick and Hinrich Schütze. 2020. Exploiting cloze-questions for few-shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676* (2020).

[45] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Erik Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2020), 4222–4235.

[46] Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of big data* 6, 1 (2019), 1–48.

[47] Michael Sutton et al. 2024. *Fuzzing: Brute Force Vulnerability Discovery.* Addison-Wesley Professional.

[48] 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.

[49] Zheng Wang, Cheng Long, and Gao Cong. 2021. Similar sports play retrieval with deep reinforcement learning. *IEEE Transactions on Knowledge and Data Engineering* 35, 4 (2021), 4253–4266.

[50] Zheng Wang, Shu Xian Teo, Jieer Ouyang, Yongjun Xu, and Wei Shi. 2024. M-RAG: Reinforcing Large Language Model Performance through Retrieval-Augmented Generation with Multiple Partitions. *ACL* (2024).

[51] Guoyang Zeng, Xie Liu, Benfeng Zhang, Yifan Zheng, Xiangwen Huang, Hongbo Chen, Jiqiang Tang, Jian Yang, and Endong Wu. 2022. Pangu-α: Large-scale autoregressive pretrained chinese language models with auto-parallel computation. *arXiv preprint arXiv:2204.12336* (2022).

[52] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2017. Generative adversarial networks: A survey and taxonomy. *arXiv preprint arXiv:1710.07035* (2017).

[53] Jia Zhang, Zheng Wang, Qian Li, Jialin Zhang, Yanyan Lan, Qiang Li, and Xiaoming Sun. 2017. Efficient delivery policy to minimize user traffic consumption in guaranteed advertising. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.

[54] Qianru Zhang, Zheng Wang, Cheng Long, and Siu-Ming Yiu. 2022. On predicting and generating a good break shot in billiards sports. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*. SIAM, 109–117.

[55] Qianru Zhang, Zheng Wang, Cheng Long, and Siu-Ming Yiu. 2024. Billiards Sports Analytics: Datasets and Tasks. *arXiv preprint arXiv:2407.19686* (2024).

[56] Michael Zimmer. 2010. "But the data is already public": on the ethics of research in Facebook. *Ethics and information technology* 12, 4 (2010), 313–325.