ALVR 2024

The 3rd Workshop on Advances in Language and Vision Research

Proceedings of the Workshop

August 16, 2024

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-153-7

Introduction

Welcome to the 3rd Workshop on Advances in Language and Vision Research. Co-located with ACL 2024, the workshop is scheduled for August 16, 2024. To facilitate the participation of the global NLP and CV community, we continue running the workshop in a hybrid format.

Language and vision research has attracted great attention from both natural language processing (NLP) and computer vision (CV) researchers. Gradually, this area is shifting from passive perception, templated language, and synthetic imagery/environments to active perception, natural language, and photo-realistic simulation or real-world deployment. The workshop covers (but is not limited to) the following topics:

- Self-supervised vision and language pre-training;
- New tasks and datasets that provide real-world solutions in language and vision;
- Text-to-image/video generation and text-guided image/video editing;
- External knowledge integration in visual and language understanding;
- Visually-grounded natural language understanding and generation;
- Language-grounded visual recognition and reasoning;
- Language-grounded embodied agents, e.g., vision-and-language navigation;
- Visually-grounded multilingual study, e.g., multimodal machine translation;
- Shortcomings of the existing large vision & language models on downstream tasks and solutions;
- Ethics and bias in large vision & language models;
- Multidisciplinary study that may involve linguistics, cognitive science, robotics, etc.;
- Explainability and interpretability in large vision & language models.

Our agenda features keynote speeches, hybrid talk sessions both for long and short papers, and poster sessions. This year we received 35 submissions, and after a thorough peer-review process, 31 papers were accepted. Among the accepted papers, 18 are archive papers and 13 are non-archive papers.

We would like to deeply thank all the authors, committee members, keynote speakers, and participants for helping us grow this research community both in quantity and quality.

Workshop Chairs

Jing Gu, UC Santa Cruz
Tsu-Jui Fu, UC Santa Barbara
Drew Hudson, Google DeepMind
Asli Celikyilmaz, Fundamentals AI Research @ Meta
William Wang, UC Santa Barbara

Organizing Committee

General Chair

Jing Gu, University of California Santa Cruz, USA
Tsu-Jui (Ray) Fu, Apple, USA
Drew Hudson, Google DeepMind, USA
Asli Celikyilmaz, Fundamentals AI Research (FAIR) @ Meta, USA
William Wang, University of California Santa Barbara, USA

Program Committee

Reviewers

Arjun Reddy Akula, Google Research
Asma Ben Abacha, Microsoft, USA
Yupeng Cao, Stevens Institute of Technology
Zhi Chen, Stevens Institute of Technology
Dhivya Chinnappa, Thomson Reuters
Zhiyang Deng, Stevens Institute of Technology
Simon Dobnik, University of Gothenburg
Yue Fan, University of California, Santa Cruz
Zhe Gan, Apple
Cristina Garbacea, University of Michigan - Ann Arbor
Ran Gong, University of California, Los Angeles
Xueying Jia, Carnegie Mellon University
Huaizu Jiang, Northeastern University
Yuechen Jiang, New Jersey Institute of Technology
Yiqiao Jin, Georgia Institute of Technology
Liqiang Jing, University of Texas at Dallas
Haohang Li, Stevens Institute of Technology
Zikai Lin, University of Michigan - Ann Arbor
Tianrui Liu, StubHub
Yujie Lu, UC Santa Barbara
Ziqian Luo, Oracle
Bohan Lyu, Tsinghua University
Loitongbam Sanayai Meetei, National Institute of Technology Silchar
Yulei Niu, ByteDance Inc.
Udita Patel, Amazon
Zhejian Peng, Snap Inc. and Georgia Institute of Technology
Vikas Raunak, Microsoft
Anthony Reyna, University of California, Santa Cruz
Arka Sadhu, University of Southern California
Shubham Shukla, Nordstrom
Thoudam Doren Singh, National Institute of Technology Silchar
Alok Singh, University of Oxford
Alessandro Suglia, Heriot-Watt University
Dachun Sun, Department of Computer Science, University of Illinois at Urbana-Champaign
Ece Takmaz, Utrecht University
Hao Tan, Adobe Systems
Yunlong Tang, University of Rochester
Jingxiao Tian, University of California, San Diego
Tianshi Wang, Department of Computer Science, University of Illinois at Urbana-Champaign
Zheng Wang, Stanford University
Yiming Xie, Northeastern University
Qi Xin, University of Pittsburgh
Wei Xiong, Adobe
Qingyang Xu, Massachusetts Institute of Technology
Qianqi Yan, University of California, Santa Cruz
Tiankai Yang, University of Southern California

Peiyan Zhang, Hong Kong University of Science and Technology
Siqiao Zhao, Morgan Stanley
Kaizhi Zheng, University of California, Santa Cruz
Chang Zhou, Columbia University
Wanrong Zhu, University of California, Santa Barbara
Fangrui Zhu, Northeastern University

Table of Contents

WISMIR3: A Multi-Modal Dataset to Challenge Text-Image Retrieval Approaches
Florian Schneider and Chris Biemann 1

mBLIP: Efficient Bootstrapping of Multilingual Vision-LLMs
Gregor Geigle, Abhay Jain, Radu Timofte and Goran Glavaš 7

LMPT: Prompt Tuning with Class-Specific Embedding Loss for Long-Tailed Multi-Label Visual Recognition
Peng Xia, Di Xu, Ming Hu, Lie Ju and Zongyuan Ge 26

Negative Object Presence Evaluation (NOPE) to Measure Object Hallucination in Vision-Language Models
Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji and Pascale Fung 37

How and where does CLIP process negation?
Vincent Quantmeyer, Pablo Mosteiro and Albert Gatt 59

Enhancing Continual Learning in Visual Question Answering with Modality-Aware Feature Distillation
Malvina Nikandrou, Georgios Pantazopoulos, Ioannis Konstas and Alessandro Suglia 73

English-to-Japanese Multimodal Machine Translation Based on Image-Text Matching of Lecture Videos
Ayu Teramen, Takumi Ohtsuka, Risa Kondo, Tomoyuki Kajiwara and Takashi Ninomiya 86

VideoCoT: A Video Chain-of-Thought Dataset with Active Annotation Tool
Yan Wang, Yawen Zeng, Jingsheng Zheng, Xiaofen Xing, Jin Xu and Xiangmin Xu 92

Enhancing Conceptual Understanding in Multimodal Contrastive Learning through Hard Negative Samples
Philipp J. Rösch, Norbert Oswald, Michaela Geierhos and Jindřich Libovický 102

Vision Language Models for Spreadsheet Understanding: Challenges and Opportunities
Shiyu Xia, Junyu Xiong, Haoyu Dong, Jianbo Zhao, Yuzhang Tian, Mengyu Zhou, Yeye He, Shi Han and Dongmei Zhang 116

SlideAVSR: A Dataset of Paper Explanation Videos for Audio-Visual Speech Recognition
Hao Wang, Shuhei Kurita, Shuichiro Shimizu and Daisuke Kawahara 129

Causal and Temporal Inference in Visual Question Generation by Utilizing Pre-trained Models
Zhanghao Hu and Frank Keller 138

Improving Vision-Language Cross-Lingual Transfer with Scheduled Unfreezing
Max Reinhardt, Gregor Geigle, Radu Timofte and Goran Glavaš 155

Automatic Layout Planning for Visually-Rich Documents with Instruction-Following Models
Wanrong Zhu, Ruiyi Zhang, Jennifer Healey, William Yang Wang and Tong Sun 167

SEA-VQA: Southeast Asian Cultural Context Dataset For Visual Question Answering
Norawit Urailertprasert, Peerat Limkonchotiwat, Supasorn Suwajanakorn and Sarana Nutanong 173

Wiki-VEL: Visual Entity Linking for Structured Data on Wikimedia Commons
Philipp Bielefeld, Jasmin Geppert, Necdet Güven, Melna Treesa John, Adrian Ziupka, Lucie-Aimée Kaffee, Russa Biswas and Gerard De Melo 186

VerbCLIP: Improving Verb Understanding in Vision-Language Models with Compositional Structures
Hadi Wazni, Kin Ian Lo and Mehrnoosh Sadrzadeh 195

Evolutionary Reward Design and Optimization with Multimodal Large Language Models
Ali Emre Narin 202

Program

Friday, August 16, 2024

09:00 - 09:15	*Welcome Speech*
09:15 - 10:15	*Invited Talk by Dr. Alane Suhr*
10:15 - 10:30	*Poster Session 1*
10:30 - 11:00	*Coffee Break*
11:00 - 12:00	*Invited Talk by Dr. Angel Chang*
12:00 - 13:00	*Invited Talk by Dr. Daniel Fried*
13:00 - 13:45	*Lunch Break*
13:45 - 14:45	*Invited Talk by Roozbeh Mottaghi*
14:45 - 15:45	*Invited Talk by Dr. Heng Ji*
15:45 - 16:15	*Coffee Break*
16:15 - 17:15	*Invited Talk by Xin (Eric) Wang*
17:15 - 17:45	*Poster Session 2*
17:45 - 18:00	*Closing Remarks*