

3rd Workshop on Advances in Language and Vision Research (ALVR 2024)

Bangkok, Thailand
16 August 2024

ISBN: 979-8-3313-0155-2

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571



Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2024) by the Association for Computational Linguistics
All rights reserved.

Printed with permission by Curran Associates, Inc. (2024)

For permission requests, please contact the Association for Computational Linguistics
at the address below.

Association for Computational Linguistics
209 N. Eighth Street
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006
Fax: 1-570-476-0860

acl@aclweb.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

<i>WISMIR3: A Multi-Modal Dataset to Challenge Text-Image Retrieval Approaches</i> Florian Schneider and Chris Biemann	1
<i>mBLIP: Efficient Bootstrapping of Multilingual Vision-LLMs</i> Gregor Geigle, Abhay Jain, Radu Timofte and Goran Glavaš	7
<i>LMPT: Prompt Tuning with Class-Specific Embedding Loss for Long-Tailed Multi-Label Visual Recognition</i> Peng Xia, Di Xu, Ming Hu, Lie Ju and Zongyuan Ge	26
<i>Negative Object Presence Evaluation (NOPE) to Measure Object Hallucination in Vision-Language Models</i> Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji and Pascale Fung	37
<i>How and where does CLIP process negation?</i> Vincent Quantmeyer, Pablo Mosteiro and Albert Gatt	59
<i>Enhancing Continual Learning in Visual Question Answering with Modality-Aware Feature Distillation</i> Malvina Nikandrou, Georgios Pantazopoulos, Ioannis Konstas and Alessandro Suglia	73
<i>English-to-Japanese Multimodal Machine Translation Based on Image-Text Matching of Lecture Videos</i> Ayu Teramen, Takumi Ohtsuka, Risa Kondo, Tomoyuki Kajiwara and Takashi Ninomiya	86
<i>VideoCoT: A Video Chain-of-Thought Dataset with Active Annotation Tool</i> Yan Wang, Yawen Zeng, Jingsheng Zheng, Xiaofen Xing, Jin Xu and Xiangmin Xu	92
<i>Enhancing Conceptual Understanding in Multimodal Contrastive Learning through Hard Negative Samples</i> Philipp J. Rösch, Norbert Oswald, Michaela Geierhos and Jindřich Libovický	102
<i>Vision Language Models for Spreadsheet Understanding: Challenges and Opportunities</i> Shiyu Xia, Junyu Xiong, Haoyu Dong, Jianbo Zhao, Yuzhang Tian, Mengyu Zhou, Yeye He, Shi Han and Dongmei Zhang	116
<i>SlideAVSR: A Dataset of Paper Explanation Videos for Audio-Visual Speech Recognition</i> Hao Wang, Shuhei Kurita, Shuichiro Shimizu and Daisuke Kawahara	129
<i>Causal and Temporal Inference in Visual Question Generation by Utilizing Pre-trained Models</i> Zhanghao Hu and Frank Keller	138
<i>Improving Vision-Language Cross-Lingual Transfer with Scheduled Unfreezing</i> Max Reinhardt, Gregor Geigle, Radu Timofte and Goran Glavaš	155
<i>Automatic Layout Planning for Visually-Rich Documents with Instruction-Following Models</i> Wanrong Zhu, Ruiyi Zhang, Jennifer Healey, William Yang Wang and Tong Sun	167
<i>SEA-VQA: Southeast Asian Cultural Context Dataset For Visual Question Answering</i> Norawit Urailetrprasert, Peerat Limkonchotiwat, Supasorn Suwajanakorn and Sarana Nutanong	173
<i>Wiki-VEL: Visual Entity Linking for Structured Data on Wikimedia Commons</i> Philipp Bielefeld, Jasmin Geppert, Necdet Güven, Melna Treasa John, Adrian Ziupka, Lucie-Aimée Kaffee, Russa Biswas and Gerard De Melo	186

<i>VerbCLIP: Improving Verb Understanding in Vision-Language Models with Compositional Structures</i>	
Hadi Wazni, Kin Ian Lo and Mehrnoosh Sadrzadeh	195
<i>Evolutionary Reward Design and Optimization with Multimodal Large Language Models</i>	
Ali Emre Narin	202