Workshop on Evaluation of Multi-Modal Generation (EvalMG 2025)

Abu Dhabi, United Arab Emirates
20 January 2025

ISBN: 979-8-3313-1365-4

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© (2025) by the Association for Computational Linguistics
All rights reserved.

Printed with permission by Curran Associates, Inc. (2026)

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

EvalMG 2025

The Workshop of Evaluation of Multi-Modal Generation

Proceedings of the The First Workshop of Evaluation of Multi-Modal Generation

January 20, 2025

©2025 The Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

Messages from the Organizers

Multimodal generation techniques have heralded new possibilities in creative content generation. Yet, the evaluation of such multimodal outputs remains a largely uncharted area, with fundamental questions still unresolved. These include understanding the contributions of individual modalities, the utility of pre-trained large language models in multimodal contexts, and the metrics for assessing faithfulness and fairness in generated outputs.

The first EvalMG workshop seeks to address these gaps by convening leading minds from natural language processing, computer vision, and multimodal AI. Our objective is to spearhead the development of robust evaluation methodologies that will propel further research in multimodal generation.

We received 21 submissions for this workshop, out of which 7 were accepted, including 5 long papers and 2 short papers. We have invited 11 reviewers and each submission was rigorously reviewed by at least three reviewers. The meta-reviews and final decisions were collaboratively handled by the organizing team.

We extend our deepest gratitude to all contributors—authors, reviewers, and particularly The University of Adelaide, for their generous support of this workshop. Your collective efforts are instrumental in shaping the future of multimodal research.

Organizing Committee

- Wei Emma Zhang
- Xiang Dai
- Desmond Elliot
- Byron Fang
- Haojie Zhuang
- Mong Yuan Sim
- Weitong Chen

- Necva Bölücü
- Danae Sanchez Villegas
- Wenhao Liang
- Lipin Guo
- Ali Shakeri
- Yutong Qu

Table of Contents

A Dataset for Programming-based Instructional Video Classification and Question Answering
Sana Javaid Raja, Adeel Zafar and Aqsa Shoaib 1

CVT5: Using Compressed Video Encoder and UMT5 for Dense Video Captioning
Mohammad Javad Pirhadi, Motahhare Mirzaei and Sauleh Eetemadi 10

If I feel smart, I will do the right thing: Combining Complementary Multimodal Information in Visual Language Models
Yuyu Bai and Sandro Pezzelle 24

LLaVA-RE: Binary Image-Text Relevancy Evaluation with Multimodal Large Language Model
Tao Sun, Oliver Liu, JinJin Li and Lan Ma 40

Persian in a Court: Benchmarking VLMs In Persian Multi-Modal Tasks
Farhan Farsi, Shahriar Shariati Motlagh, Shayan Bali, Sadra Sabouri and Saeedeh Momtazi 52

TaiwanVQA: A Benchmark for Visual Question Answering for Taiwanese Daily Life
Hsin-Yi Hsieh, Shang Wei Liu, Chang Chih Meng, Shuo-Yueh Lin, Chen Chien-Hua, Hung-Ju Lin, Hen-Hsen Huang and I-Chen Wu 57

Guiding Vision-Language Model Selection for Visual Question-Answering Across Tasks, Domains, and Knowledge Types
Neelabh Sinha, Vinija Jain and Aman Chadha 76