

ALVR 2026

**Workshop on Advances in Language and Vision Research**

**Proceedings of the Workshop**

July 3, 2026

The ALVR organizers gratefully acknowledge the support from the following sponsors.

## **Gold**

 **Lambda**

©2026 Association for Computational Linguistics

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](mailto:acl@aclweb.org)

ISBN 979-8-89176-398-2

## Introduction

We are excited to welcome you to the 4th Workshop on Advances in Language and Vision Research (ALVR), co-located with ACL 2026 and held on July 3, 2026, in San Diego, California.

ALVR provides a dedicated forum for research at the intersection of natural language processing and computer vision. This year’s program features five invited talks from leading researchers, six spotlight presentations, and twenty-three poster presentations, spanning a wide range of topics at the frontier of language and vision research.

We received 47 submissions this year. After a thorough review process, we accepted 29 papers in total: 23 as archival papers and 6 through our non-archival track, which welcomes work concurrently under submission or recently published elsewhere. The archival acceptance rate is approximately 49%. Each submission received at least three reviews, with area chairs providing guidance to ensure the quality and breadth of the final program. We thank our area chairs—Yue Fan, Syrielle Montariol, Alane Suhr, Xin Eric Wang, and Qianqi Yan—for their careful deliberation, and all program committee members for their thorough and timely reviews.

We are grateful to our five distinguished invited speakers: Raymond J. Mooney (University of Texas at Austin), Lianhui Qin (UC San Diego), Amir Zadeh (Lambda Labs), Mohit Bansal (University of North Carolina at Chapel Hill), and Jiajun Wu (Stanford University), for sharing their insights and inspiring the community.

We gratefully acknowledge the generous support of our sponsor, Lambda Labs (Gold level), whose contribution has been instrumental to the success of the workshop.

We extend our sincere thanks to all authors who submitted to ALVR 2026, and to all attendees for their enthusiasm and participation. It is the community’s energy and commitment that makes this workshop a rewarding venue for advancing research at the intersection of language and vision.

Qianqi Yan, Lead Organizer

Syrielle Montariol, Yue Fan, Jing Gu, Jiayi Pan, Manling Li, Parisa Kordjamshidi, Alane Suhr, and Xin Eric Wang, Co-Organizers

# Organizing Committee

## Organizers

Qianqi Yan, University of California, Santa Barbara, USA  
Syrielle Montariol, University of California, Berkeley, USA  
Yue Fan, Autodesk Research, USA  
Jing Gu, xAI, USA  
Jiayi Pan, xAI, USA  
Manling Li, Northwestern University, USA  
Parisa Kordjamshidi, Michigan State University, USA  
Alane Suhr, University of California, Berkeley, USA  
Xin Eric Wang, University of California, Santa Barbara, USA

## Program Committee

### Area Chairs

Qianqi Yan, University of California, Santa Barbara  
Yue Fan, Autodesk  
Syrielle Montariol, ISIR, Sorbonne Université, France  
Alane Suhr, University of California, Berkeley  
Xin Eric Wang, University of California, Santa Barbara and Simular

### Reviewers

Rishabh Agrawal, Amazon  
Yifei Bi, Facebook  
Zhong Chen, Southern Illinois University-Carbondale  
Sahil Rajesh Dhayalkar, Brain Corporation  
Christopher Driggers-Ellis, University of Florida  
Yue Fan, Autodesk  
Zhiyuan Gao, University of Southern California  
Jingnan Gao, NVIDIA and Shanghai Jiao Tong University  
Christan Grant, University of Florida  
Jinru Han, University of California, Los Angeles  
Patrick Huber, Facebook  
Ben Jenkins, Florida Atlantic University  
Yanru Jiang, University of California, Los Angeles  
Dayeon Ki, Microsoft and University of Maryland, College Park  
Sriram Kollipara, Walmart  
Pawan Kumar, International Institute of Information Technology Hyderabad  
Yuecheng Li, Kuaishou  
Chuhan Li, University of California, Santa Barbara  
Irene Li, University of Tokyo  
Mao Lin, University of California, Merced  
Shih-chih Lin, National Tsinghua University  
Junbin Lu, University of Washington  
Sepideh Mamooler, EPFL  
Srijith Ravikumar, Amazon  
Anisha Saha, Saarland Informatics Campus, Max-Planck Institute  
Tongyue Shi, Peking University  
Andrew Shin, Keio University  
Kaleen Shrestha, University of Southern California  
Thoudam Doren Singh, National Institute of Technology Meghalaya  
Sakthivel Sivaraman, NVIDIA  
Elior Sulem, Ben-Gurion University of the Negev  
Zhimin Sun, AMS AI Lab, Tencent  
Rohith Uppala, LinkedIn  
Pavan Kumar Velaga, Amazon and Rutgers University  
Zirui Wei, C3.ai  
Yibo Yan, HKUST and Alibaba  
Yang Yan, Southern Illinois University-Carbondale  
Junhuan Yang, Amazon

Rongtian Ye, Aalto University  
Ziyao Zeng, Yale University  
Jing Zhang, Amazon  
Pingyue Zhang, Northwestern University  
Hang Zhao, Northeastern University  
Duo Zhou, University of Illinois at Urbana-Champaign

## Table of Contents

<i>Thinking in Pictures: A Diagnostic Study of Visual vs. Textual Chain-of-Thought Reasoning in Vision-Language Models</i> Ben Jenkins .....	1
<i>A Zipfian Analysis of Visual Token Distributions for AI-Generated Images</i> Andrew Shin .....	13
<i>Semantically Aware Optimal Transport for Dense Label Transfer</i> Preeti, Kiran Ravish, Ankita Kushwaha and Pawan Kumar .....	18
<i>CoSMoEs: Compact Sparse Mixture of Experts</i> Patrick Huber, Akshat Shrivastava, Ernie Chang, Chinnadhurai Sankar, Ahmed A Aly and Adithya Sagar .....	46
<i>GraphicWeaver: Benchmarking Agentic Planning for Graphic Design Generation</i> Dayeon Ki, Tianyi Zhou, Marine Carpuat, Gang Wu, Puneet Mathur and Viswanathan Swaminathan .....	57
<i>Scaling Vision–Language Models for Pharmaceutical Long-Form Video Reasoning on Industrial GenAI Platform</i> Suyash Mishra, Qiang Li, Srikanth Patil, Satyanarayan Pati and Baddu Narendra .....	85
<i>PGGA: A Plan-Grounded GUI Agent for Automated Device Support</i> Lei Hsiung, Zhiyu Chen, Seonhoon Kim and Qun Liu .....	105
<i>CAFES: A Collaborative Multi-Agent Framework for Multi-Granular Multimodal Essay Scoring</i> Jiamin Su, Yibo Yan, Zhuoran Gao, Han Zhang, Xiang Liu, Huiyu Zhou and Xuming Hu ...	115
<i>GM-PRM: A Generative Multimodal Process Reward Model for Multimodal Mathematical Reasoning</i> Jianghangfan Zhang, Yibo Yan, Kening Zheng, Xin Zou, Song Dai and Xuming Hu .....	139
<i>Look Where You’re Told: Instruction-Consistent Attention for GUI Grounding</i> Seonhoon Kim, Zhiyu Chen, Xin Li and Qun Liu .....	155
<i>From Pixels to BFS: High Maze Accuracy Does Not Imply Visual Planning</i> Alberto Gonzalo Rodriguez Salgado .....	164
<i>When Relations Break: Analyzing Relation Hallucination in Vision-Language Model Under Rotation and Noise</i> Philip Wootae Shin, Ajay Narayanan Sridhar, Lakshmi Sivani Devarapalli, Rui Zhang, Jack Sampson and Vijaykrishnan Narayanan .....	180
<i>VLCE: A Knowledge-Enhanced Framework for Image Description in Disaster Assessment</i> Md. Mahfuzur Rahman, Marufa Kamal, Fahad Rahman, Sunzida Siddique, Ahmed Rafi Hasan, Mohd Ariful Haque, Kishor Datta Gupta and Roy George .....	186
<i>Beyond Visual Similarity: Rule-Guided Multimodal Clustering with explicit domain rules</i> Kishor Datta Gupta, Mohd Ariful Haque, Marufa Kamal, Ahmed Rafi Hasan, Md. Mahfuzur Rahman and Roy George .....	199
<i>ChartDiff: A Large-Scale Benchmark for Comprehending Pairs of Charts</i> Rongtian Ye .....	209

<i>Formal Machine Interpretation for the Semasiographic Mixtec Codices of Precolonial and Early Colonial Mesoamerica</i>	
Christopher Driggers-Ellis, Gabriel Ayoubi, girish.salunke811@gmail.com girish.salunke811@gmail.com and Christan Grant .....	230
<i>Temporal-Linguistic Adaptive Streaming for Continuous Sign Language Translation</i>	
Arshia Kermani, Habib Irani, Deautun Ross and Vangelis Metsis .....	239
<i>FADE: Probing the Limits of VLMs on fine-grained OCR</i>	
Deep Shah, Nehal Kathrotia and Sanket Badhe .....	249
<i>Efficient Visual Grounding in VQA via Question-Guided Sparse Attention</i>	
Prasanth .....	260
<i>Systematic Performance Degradation in Indic Vision-Language Models: Evidence from Hindi and Telugu</i>	
Rishikant Chigrupaatii, Ponnada Sai Tulasi Kanishka, Lalit Chandra Routhu, Martin Patel, Sama Supratheek Reddy, Divyam Gupta, Rajiv Misra and Rohun Tripathi .....	272
<i>How Fragile Is Vision-Language Alignment? Mapping Concept Disruption Under Text-to-Image Personalization</i>	
Mujtaba Hasan .....	278
<i>The Compositional Grounding Gap: Why Vision-Language Models Fail at Relational Reasoning and How to Fix It</i>	
Kaustubh S. Bukkapatnam .....	287
<i>HalluTrace: Causal Attribution and Source-Targeted Decoding for Hallucination in Large Vision-Language Models</i>	
Kaustubh S. Bukkapatnam .....	294