MAGMaR 2026

The 2nd Workshop on Multimodal Augmented Generation via Multimodal Retrieval

Proceedings of the Workshop

July 4, 2026

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 979-8-89176-425-5

Introduction

We are delighted to welcome you to MAGMaR 2026, the second workshop on Multimodal Augmented Generation via Multimodal Retrieval. MAGMaR is being held in San Diego, USA on the 4th of July, 2026, and is co-located with ACL 2026, which takes place from July 2nd through the 7th. This workshop is organized in support of ACL's Special Interest Group on Image and Language (SIGIL).

Audiovisual media is becoming an increasingly dominant form of online information consumption. From firsthand, "in the wild" video footage of natural disasters to professionally edited news coverage of major political events, videos serve as rich sources of information for producing factual, grounded articles. Especially for actively unfolding events, grounding articles in video can help combat misinformation and provide journalists and analysts with tools to quickly synthesize new developments.

Individual research groups have independently begun addressing this challenge, leading to parallel yet disconnected efforts to define the research space. ACL 2025 hosted the first MAGMaR workshop focused on Video Event Retrieval. This year's iteration focused on two primary areas: (1) the retrieval of multimodal content spanning text, images, audio, and video; and (2) retrieval-augmented generation, with an emphasis on multimodal retrieval and grounded generation. Relevant topics to the workshop this year included document retrieval, multimodal retrieval, retrieval-augmented generation (RAG), multimodal RAG, multimodal question answering, and research on video, image, and audio understanding.

To further this goal, we again hosted a shared task focused on video retrieval, and moreover, extended the task this year to include article generation from multiple videos. Specifically, it focused on retrieving relevant videos and generating grounded reports that respond to information needs. Given a query describing a real-world current event, participating systems needed to identify pertinent videos from a large multilingual, multimodal collection and use that evidence to produce a coherent and informative written report.

There were two tracks:

- Retrieval: Systems provided a ranked list of videos in the collection ordered by relevance to the query.
- Generation: Systems produced a text report that answers the information need and grounds its content in the retrieved videos. Teams were able to submit to either track or both.

We saw a large increase in the number of submissions to our shared task this year, with four teams submitting dozens of systems. All teams had at least one system that beat a very strong baseline and yielded some very interesting insights on what works and where are the open problems in this challenging multimodal domain. Check out the findings paper and teams' system descriptions for some really interesting analysis of how to build strong Video RAG systems.

This year, the program of MAGMaR includes two keynote talks, one presentation session, and one poster session. With an increse in submissions from last year, we were able to accept 15 out of 26 papers, for an overall acceptance rate of 58%. Of these, six were accepted as oral presentations. Once more, we allowed for non-archival submissions which has led to some interesting papers published in other venues that are being presented at the workshop. The members of our Program Committee and Organizing Comittee did an excellent job in reviewing the submitted papers, and we thank them for their essential role in selecting the accepted papers and helping produce a high quality program for the conference.

A workshop requires the hard work of numerous people, both behind the scenes and those that you will see more prominently. First off, we want to say thank you to our two keynote speakers, Nanyun (Violet) Peng from UCLA and Chenliang Xu from the University of Rochester, who have agreed to give talks about multimodal problems. Dr. Peng's talk "Towards Self-Improving Multimodal Models" tackles multimodal reasoning and generation problems, while Dr. Xu's talk "Multi-level Alignment in Audio-Visual Scene Generation and Learning" looks at aligning representations across modalities. Both of these cover challenging problems focused on in this workshop and explored in our shared task. We appreciate their insights.

Additonally, we would be remiss to not mention the people who helped organize (and participated) in our shared task on retrieving events in videos. Our online leaderboard received numerous submissions and grew substantially over last year.

Finally, we thank all contributors, reviewers, and attendees who helped make MAGMaR 2026 possible. We hope you enjoy a day full of engaging talks, thought-provoking posters, and stimulating discussion.

Kenton Murray and Reno Kriz, Editors

Organizing Committee

Organizers

Kenton Murray, Human Language Technology Center of Excellence, Johns Hopkins University
Reno Kriz, Human Language Technology Center of Excellence, Johns Hopkins University
Andrew Yates, Human Language Technology Center of Excellence, Johns Hopkins University
Francis Ferraro, University of Maryland, Baltimore County
Desmond Elliott, University of Copenhagen
Xiang Xiang, Huazhong University of Science and Technology
Alexander Martin, Johns Hopkins University
Joel Brogan, OpenAI
Teng Long, University of Amsterdam, University of Trento
Jeremy Gwinnup, Air Force Research Labratory

Program Committee

Program Committee

Xiang Xiang, Huazhong University of Science and Technology
Dengjia Zhang, Johns Hopkins University
Reno Kriz, Human Language Technology Center of Excellence, Johns Hopkins University
Kenton Murray, Human Language Technology Center of Excellence, Johns Hopkins University
Eugene Yang, Human Language Technology Center of Excellence, Johns Hopkins University
Francis Ferraro, University of Maryland, Baltimore County
Jeremy Gwinnup, Air Force Research Laboratory
Saket Saurabh, OpenAI
Maitrik Patel, Apple
Cameron Carpenter, Johns Hopkins University
Will Walden, Human Language Technology Center of Excellence, Johns Hopkins University
David Etter, Human Language Technology Center of Excellence
Andrew Yates, Human Language Technology Center of Excellence, Johns Hopkins University
Tyler Skow, Johns Hopkins University
Alexnder Martin, Johns Hopkins University
Goonjan Saha, Samsung
Parin Rajesh Jhaveri, J.P. Morgan Chase
Sahil Rajesh Dhayalkar, Brain Corporation
Joel Brogan, OpenAI
Teng Long, University of Amsterdam
Tejas Gokhale, University of Maryland, Baltimore County
Debashish Chakraborty, Human Language Technology Center of Excellence, Johns Hopkins University
Desmond Elliott, University of Copenhagen

Invited Speakers

Nanyun (Violet) Peng, University of California Los Angeles
Chenliang Xu, University of Rochester

Keynote Talk

Towards Self-Improving Multimodal Models

Dr. Nanyun (Violet) Peng
Associate Professor
Department of Computer Science
University of California Los Angeles
2026-07-04 09:45:00 – Room: **Old Town**

Abstract: Large multimodal models (LMMs) have made impressive progresses and performed well on tasks such as image captioning, visual question answering, and grounded dialogue. Yet despite this progress, they continue to struggle with two fundamental challenges: learning new concepts beyond their training data, and reliably solving complex multimodal reasoning and generation tasks. Overcoming these limitations is essential if we want LMMs to function robustly in open-world settings, where new categories constantly emerge, and to support high-stakes applications like scientific analysis, education, or healthcare, which demand precise reasoning and generation. A crucial ingredient for overcoming these challenges is enabling models to reflect on and improve their own outputs. In this talk, I present three complementary efforts towards this vision. First, we explore how contrastive augmentation can expand models' conceptual coverage, helping them recognize rare or fine-grained visual categories. Second, we introduce a multi-agent framework that decomposes complex multimodal generation into specialized roles, showing how structured collaboration improves reliability and scales with computational budget. Finally, we investigate fine-grained critique and correction in visual reasoning, proposing benchmarks and strategies that highlight reflection as a pathway to better learning and reasoning. Taken together, these directions sketch a roadmap towards self-improving multimodal models –systems that can adapt, reflect, and refine themselves in pursuit of deeper visual understanding and reasoning.

Bio: Nanyun (Violet) Peng is an Associate Professor of Computer Science at the University of California, Los Angeles, currently on sabbatical, and a Senior Staff Research Scientist at Google. Her research focuses on controllable and creative generation, multilingual and multimodal models, and automatic evaluation of AI agents, with a strong commitment to advancing robust and trustworthy artificial intelligence (AI). Her work has been recognized with multiple paper awards, including an Outstanding Paper Award at NAACL 2022, three Outstanding Paper Awards at EMNLP 2024, Oral Papers at NeurIPS 2022 and ICML 2023, as well as several Best Paper Awards at workshops. Her research has received support from the NSF CAREER Award, NIH R01, DARPA, IARPA, and multiple industrial research awards. She served as Program Chair for ICLR 2025 and EMNLP 2025, and as a board member of NAACL.

Keynote Talk

Multi-level Alignment in Audio-Visual Scene Generation and Learning

Dr. Chenliang Xu
Associate Professor
Department of Computer Science
University of Rochester
2026-07-04 16:00:00 – Room: **Old Town**

Abstract: In this talk, I will discuss how to align audio, visual, spatial, and semantic representations across multiple levels, from low-level perceptual correspondence to object/event-level structure and scene-level generation. The talk connects audio-visual learning with scene understanding, generative modeling, and multimodal AI.

Bio: Chenliang Xu is a tenured Associate Professor of Computer Science at the University of Rochester. His research lies at the intersection of computer vision, audio-visual learning, and trustworthy AI, with a focus on teaching machines to understand the world through video, sound, and language. He has published over 130 papers at top venues including CVPR, NeurIPS, ICCV, ECCV, ICLR, and ICML, with support from agencies such as DARPA, NSF, and NIH. His work has received multiple best paper awards, and he has served as an area chair for major conferences in computer vision and machine learning.

Table of Contents

Program

Saturday, July 4, 2026

09:30 - 09:45 *Welcome Remarks*

09:45 - 10:30 *Keynote 1 Nanyun (Violet) Peng, UCLA*

10:30 - 11:00 *Break*

11:00 - 12:30 *Oral Presentations*

12:30 - 14:00 *Lunch*

14:00 - 15:30 *Poster Session*

15:30 - 16:00 *Break*

16:00 - 16:45 *Keynote 2 Chenliang Xu, University of Rochester*

16:45 - 17:00 *Paper Awards and Closing*