MAGMaR 2025

Workshop on Multimodal Augmented Generation via Multimodal Retrieval

Proceedings of the Workshop

August 1, 2025

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 979-8-89176-280-0

Introduction

We are excited to welcome you to MAGMaR 2025, the first workshop on Multimodal Augmented Generation via Multimodal Retrieval. MAGMaR is being held in Vienna, Austria on August 1, 2025, and is co-located with ACL 2025, which takes place from July 28th-August 1st.

While information retrieval systems for text documents have been extensively studied for decades, the landscape has shifted; vast amounts of information today are stored as videos with minimal text metadata. For instance, online commercial platforms host billions videos. Despite the explosion of multimodal data, there remains a dearth of research around the efficient retrieval, processing, and synthesis of these massive multimodal collections. Existing systems largely still rely on text metadata (e.g., human written descriptions), overlooking the rich semantic content embedded within the multimodal data itself.

Individual research groups have independently begun addressing this challenge, leading to parallel yet disconnected efforts to define the research space. MAGMaR was conceived as a collaborative venue to unify these efforts and foster dialogue, which we believe is crucial for advancing the field. The MAGMaR workshop focuses on two primary areas: (1) the retrieval of multimodal content, which spans text, images, audio, video, and multimodal data (e.g., image-language, video-language); and (2) retrieval-augmented generation, with an emphasis on multimodal retrieval and generation.

To further this goal, we hosted a shared task on event-based video retrieval and understanding, designed to spark interest and facilitate research development in both retrieval and generation. This task's primary retrieval metric, nDCG@10, compared the final ranked lists of videos produced by participant systems.

The shared task was built around MultiVENT 2.0 (Kriz et al., 2024). While prior datasets like MSR-VTT (10,000 videos) and MultiVENT (2,400 videos; Sanders et al., 2023) made progress toward multilingual and event-centric video retrieval, they remain small compared to typical text retrieval corpora—e.g., HC4 from the 2022 NeuCLIR shared task, which contains 6 million documents. To address this gap, we introduced MultiVENT 2.0, a large-scale benchmark with over 217,000 videos and 2,549 event-centric queries for a test collection of 109,800 videos. The dataset covers a diverse range of real-world current events and is designed to facilitate both retrieval and generation research.

MultiVENT 2.0 has been made publicly available on HuggingFace[1] and includes extracted features such as visual frames, transcribed speech, embedded text, and frame-level captions. Relevance judgments for the training set were released publicly, while those for the test set were hosted on an Eval.ai leaderboard[2]. The primary task setting restricts participants to use only the raw video content; using additional metadata or text descriptions is permitted only in an oracle setting.

Several teams submitted strong systems to the leaderboard. The best-performing submission, OmniEmbed, was developed by the Tevatron group from the University of Waterloo: Jiaqi Samantha Zhan, Crystina Zhang, Shengyao Zhuang, Xueguang Ma, and Jimmy Lin. Their best non-oracle system achieved an nDCG@10 of 0.709, a significant improvement over the strongest original baseline (0.324).

This year, the program of MAGMaR includes two keynote talks, one presentation session, and one poster session. In our inaugural year, we received 21 submissions and accepted 14, for an overall acceptance rate of 67%. Of these, five were accepted as oral presentations. 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.

[1] https://huggingface.co/datasets/hltcoe/MultiVENT2.0

[2] https://eval.ai/web/challenges/challenge-page/2507

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, Desmond Elliott (University of Copenhagen) and Joel Brogan (Oak Ridge National Laboratory) whose interdisciplinary talks are a nice resource for the broader NLP and ACL communities. Both Dr. Elliott's talk "Recent Experiments in Retrieval-Augmented Image Captionin" and Dr. Brogan's talk "When you Don't Quite Know What You Want: Bridging the Multimodal Search Intention Gap" cover challenging, state-of-the-art problems at the unique intersection of the focus of MAGMaR and we appreciate the insights that they are sharing. 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 we continue to have people engaging with it even though the official evaluation is closed.

Finally, we thank all contributors, reviewers, and attendees who helped make MAGMaR 2025 possible. We hope you enjoy a day full of engaging talks, thought-provoking posters, and stimulating discussion.

Reno Kriz and Kenton Murray, Editors

Organizing Committee

Organizers

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
Kate Sanders, Johns Hopkins University
Cameron Carpenter, Johns Hopkins University
Benjamin Van Durme, Johns Hopkins University and Microsoft

Program Committee

Program Committee

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
Kate Sanders, Johns Hopkins University
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
Alex Martin, Johns Hopkins University
Gaurav Kumar, University of California San Diego

Invited Speakers

Joel Brogan, Oak Ridge National Laboratory
Desmond Elliot, University of Copenhagen

Keynote Talk

Recent Experiments in Retrieval-Augmented Image Captioning

Dr. Desmond Elliott
Associate Professor
Department of Computer Science
University of Copenhagen

2025-08-01 09:45:00 – Room: **2.44**

Abstract: Retrieval-augmentation has proven useful in a wide-range of classification and generation tasks, and it is now powering the next generation of Large Language Models. In this talk, I will present recent research on applying retrieval-augmentation to image caption generation. I will start by outlining how retrieval-augmentation can work in this task, and present a parameter-efficient image captioning model that can describe images from a variety of domains by hot-swapping the contents in the retrieval data store without retraining the model. Then I will describe two approaches to multilingual image captioning: one based on prompting an LLM without any training, the other based on supervised training with either multilingual or monolingual data. Finally, I will speak about our efforts to understand and explain the success and failure modes of retrieval-augmented image captioning

Bio: Dr. Desmond Elliot is an Associate Professor and a Villum Young Investigator at the University of Copenhagen. His main research interests are tokenization-free language modelling, and multilingual and multimodal processing. Dr. Elliot's work received a Best Paper Honorable Mention at the CVPR 2025 Workshop on Visual Concepts, the Best Long Paper Award at EMNLP 2021, and an Area Chair Favourite paper at COLING 2018. His research is funded by the Velux Foundations, the Innovation Foundation Denmark, the Novo Nordisk Foundation, the Poul de Jensen Foundation, Meta, and Google.

Keynote Talk

When you Don't Quite Know What You Want: Bridging the Multimodal Search Intention Gap

Dr. Joel Brogan
Research Group Lead – Multimodal Sensor Analytics
Center for AI Security Research
Energy Systems and Technology Directorate
Oak Ridge National Laboratory, Department of Energy
2025-08-01 16:00:00 – Room: **2.44**

Abstract: In research and analysis, the most valuable insights often lie beyond what we think to look for. Yet building systems that can surface these unknown unknowns remains a fundamental challenge. How do you design retrieval methods for discoveries you can't define upfront, and how do you measure success when you didn't know what you wanted in the first place? In this talk, I will share some of the practical ways our team at the Multimodal Sensor Analytics Group has approached this problem. We will explore how multimodal retrieval, combining vector stores and graph-based approaches, can bridge the gap between what you are searching for and what you truly need to find. I will discuss examples where these systems have surfaced unexpected but meaningful patterns, and reflect on the limitations, opportunities, and design choices when aiming to build retrieval systems that broaden rather than narrow human attention.

Bio: Dr. Joel Brogan is a Research Professional and Group Lead of the Multimodal Sensing Analytics Group at Oak Ridge National Laboratory, a US DOE national lab. There, he leads a team of 13 researchers who perform work in inverse imaging, graph analytics, biometrics, and adversarial AI vulnerability mitigation. Dr. Brogan received his PhD in computer vision at the University of Notre Dame, where he worked under the DARPA MediFor program to design image and video retrieval and analysis algorithms to help detect and understand the dynamics of misinformation spread. He joined Oak Ridge National Laboratory in 2019, where he is currently the Evaluation Lead for the IARPA BRIAR Program, Biometric Recognition and Identification at Altitude and Range, which aims to perform large-scale biometric characterization human action from video at long distances and altitudes. Additionally, Dr. Brogan is a founding member of the Center for AI Security Research, or CAISER, through which he and his design content retrieval tools that aim to discover previously unknown patterns in large pools of multimodal data. His work has been nominated for the 2023 R&D100 awards and the AFCEA 2023 FedID Best Operational Success Award. In 2024, Dr. Brogan was Honored as a Finalist in the FedScoop 50 "Most Inspiring Up & Comer" category.

Table of Contents

Program

Friday, August 1, 2025

09:30 - 09:45 *Opening Remarks*

09:45 - 10:30 *Keynote 1 (Desmond Elliott, PhD)*

10:30 - 11:00 *Break*

11:00 - 12:30 *Oral Presentations*

MT2ST: Adaptive Multi-Task to Single-Task Learning
Siyuan Yuan

One Pic is All it Takes: Poisoning Visual Document Retrieval Augmented Generation with a Single Image (Short Paper)
Ezzeldin Shereen, Dan Ristea, Burak Hasircioglu, Shae McFadden, Vasilios Mavroudis and Chris Hicks

M2IV: Towards Efficient and Fine-grained Multimodal In-Context Learning in Large Vision-Language Models
Yanshu Li, Yi Cao, Xi Xiao and Tianyang Wang

Q2E: Query-to-Event Decomposition for Zero-Shot Multilingual Text-to-Video Retrieval
Shubhashis Roy Dipta and Francis Ferraro

OmniEmbed-MultiVent: Unified End-to-end All-modality Retrieval
Jiaqi Samantha Zhan, Crystina Zhang, Shengyao Zhuang, Xueguang Ma and Jimmy Lin

12:30 - 14:00 *Lunch*

14:00 - 15:30 *Poster Session*

15:30 - 16:00 *Break*

16:00 - 16:45 *Keynote 2 (Joel Brogan, PhD)*

16:45 - 17:30 *Best Paper Award and Closing*

Friday, August 1, 2025 (continued)