Call For Papers

SPECIAL SESSION

Large Models for Embodied Multimedia Processing

Recent advances in large-scale foundation models, including large language models (LLMs), vision-language models (VLMs), and multimodal foundation models, have significantly reshaped the landscape of multimedia processing. Beyond traditional perception and understanding tasks, these models are increasingly being integrated into embodied systems, such as robots, autonomous agents, mixed reality devices, and intelligent environments, where perception, reasoning, and action are tightly coupled.

Embodied multimedia processing introduces new challenges that go beyond static multimedia analysis. These challenges include multimodal sensory fusion (vision, audio, tactile, and proprioception), long-horizon temporal reasoning, real-time interaction, physical world grounding, and safety-critical decision-making. Large models offer unprecedented opportunities to address these challenges by providing unified representations, cross-modal reasoning capabilities, and scalable learning paradigms.

This special session aims to bring together researchers and practitioners from multimedia, robotics, computer vision, natural language processing, and embodied AI communities to explore how large models can be designed, adapted, and deployed for embodied multimedia processing. More importantly, it will foster interdisciplinary discussions on novel algorithms, system architectures, datasets, evaluation protocols, and real-world applications.

The special session invites submissions addressing, but not limited to, the following areas:

- Multimodal Foundation Models for Multimedia: Design and adaptation of large-scale multimodal models for understanding, processing, and generation of multimedia content.
- Multimedia Representation and Cross-Modal Perception: Learning robust and aligned representations across visual, auditory, textual, and other multimedia modalities.
- Spatial-Temporal Multimedia Understanding: Modeling dynamic scenes and events through spatial-temporal analysis of multimodal data.

- 3D/4D Multimedia Perception: Reconstruction and understanding of 3D/4D scenes from multimodal multimedia inputs.
- Multimedia Grounding and Embodied Perception: Grounding multimedia content in physical or interactive environments for perception-driven tasks.
- Multimedia Reasoning and Semantic Understanding: High-level semantic understanding and reasoning over multimodal multimedia signals.
- Multimedia Content Generation and Enhancement: Generation, editing, and enhancement of multimedia content using large models.
- Efficient Multimedia Processing Systems: Scalable and real-time systems for multimedia perception using large models on edge or distributed platforms.
- Multimedia Datasets and Evaluation: Benchmark datasets, evaluation protocols, and metrics for multimodal perception and multimedia understanding.

IMPORTANT DATES

- Paper submission deadline: May 31, 2026
- Notification of acceptance: June 15, 2026
- Camera-ready submission: June 30, 2026

All deadlines are at the end of the day specified, anywhere on Earth (UTC-12).

SUBMISSION INSTRUCTION

Please use the below link to submit your work.

[Please add the submission button/URL here]

Please note: Submissions to this special session must follow the same formatting guidelines, templates, page limits, and review policies as the "Regular Paper Track" of the main conference. Authors are encouraged to refer to https://mipr2026.org/authors/ for detailed instructions.

CONTACT

Dr. Song Wang, Zhengzhou University, China (ieswang[at]zzu.edu.cn)

Dr. Deyin Liu, University of Surrey, UK

Leihan Chen, Toronto Metropolitan University, Canada

Dr. Xin Guo, Zhengzhou University, China