SaFeMM-AI: Workshop on Safe and Trustworthy Multimodal AI Systems

Call for Papers

The ICCV 2025 Workshop on Safe and Trustworthy Multimodal AI Systems (SaFeMM-AI) invites submissions of original research papers focused on enhancing the safety, robustness, and reliability of multimodal models. We welcome works that propose novel methods for mitigating multimodal hallucinations, safeguarding user privacy, and defending against adversarial or jailbreak attacks. We also encourage submissions that address bias, fairness, ethical transparency, and interpretability, as well as new evaluation protocols or benchmarks for assessing safety and trustworthiness in multimodal large language models (MLLMs) and agentic systems.

Topics of interest

Our workshop focuses on advancing the development of multimodal AI systems that can robustly handle unsafe or adversarial inputs and consistently generate safe, reliable, and trustworthy outputs. Topics of interest include but are not limited to:

- Hallucination detection and mitigation in Multimodal Large Language Models (MLLMs) and agents
- Privacy-preserving methods, data protection, and copyright protection in multimodal systems
- Reliability in multimodal systems, including cross-modal consistency, controllability, and stability
- Defenses against adversarial, backdoor, and jailbreak attacks in multimodal systems
- Detection and prevention of harmful, toxic, or misleading content (e.g., deepfakes)
- Prevention of misuse in multimodal systems, including copyright infringement and plagiarism detection

- Safety and trustworthiness evaluation protocols and benchmarks for MLLMs and agents
- Bias, fairness, and ethical considerations in multimodal systems
- Safety challenges in emerging learning paradigms, such as modality bridging and test-time scaling
- Unique safety challenges in multimodal systems compared to single-modality models
- Interpretability and explainability in cross-modal reasoning and outputs
- Trustworthy agentic behavior in vision-language and multimodal AI systems
- Frameworks for responsible deployment and auditing of multimodal systems and agents
- Safe usage and risk assessment of MLLMs and agents in scientific domains (e.g., healthcare, chemistry, biology, robotics, autonomous driving)

Submission Guidelines

Paper formatting: Submitted papers must be formatted using the ICCV 2025 Author Kit and are limited to **eight pages**, including figures and tables. Additional pages are allowed only for references.

Important: Papers that exceed the page limit (excluding references), are not properly anonymized, or do not use the official ICCV template **will be rejected without review**. We **strongly encourage** authors to carefully follow the ICCV Author Guidelines, as our workshop will adhere to the same formatting and submission policies as the main conference.

Submission and review process: The review process will be double-blind, and submissions will be managed via OpenReview. During the review period, submissions will be visible only to their assigned reviewers and area chairs. Reviews and author responses will never be made public.

Anyone who plans to submit a paper as an author or a co-author will need to create (or update) their OpenReview profile by the paper submission deadline. We recommend using Institution emails to register your account on OpenReview, as non-institutional emails can take up to two weeks to register. By submitting a paper to this workshop, the authors agree to the review process and understand that papers are processed by the OpenReview system to match each manuscript to the best possible area chairs and reviewers.

OpenReview author instructions can be found here.

Submission Process

All submissions should be made through the workshop's OpenReview portal. (Submission will open soon)

Important Dates and Deadlines (Tentative)

	Deadline
Paper Submission	June 19, 2025 (23:59 AoE)
Paper Notification	July 11, 2025
Camera-ready Paper	August 11, 2025
Workshop date	October 19-20, 2025

Contact Information

For any questions, please contact us at: safemm.ai.workshop@gmail.com
For more information, visit our website: https://safemmai.github.io/