Workshop on Multimodal Agents for ECCV 2024

Zane Durante[1], Ehsan Adeli[1], Juan Carlos Niebles[1,2], Naoki Wake[3], Bidipta Sarkar[1], Ran Gong[4], Jae Sung Park[5], Yejin Choi[5], Fei-Fei Li[1], and Qiuyuan Huang[3]

[1] Stanford University
[2] Salesforce AI Research
[3] Microsoft Research, Redmond
[4] University of California, Los Angeles
[5] University of Washington

Keywords: Agent, Agent AI, Multimodal Agents, Foundation Models

Abstract. The field of artificial intelligence is experiencing a surge in the effectiveness of multimodal agents, signaling their potential to become ubiquitous in our everyday lives. Our proposed workshop is dedicated to fostering the advancement of multimodal foundation models by focusing on their development as agents in distinct environments. Our workshop strives to extend traditional agent frameworks of next-action prediction and reinforcement learning by emphasizing critical elements such as the integration of natural language, dynamic multimodal interactions, and interactive feedback with humans and the environment. In an effort to further advance research on agent-based multimodal intelligence, we propose our new workshop on multimodal agents. Our workshop is designed to create a platform for publishing research on the multifaceted aspects of agentic, multimodal interactions.

1 Summary

Acronym	MMA
Workshop Title	Workshop on Multimodal Agents
Primary organizer name and email	Zane Durante (durante@stanford.edu)
Half or full day	Half Day
Expected number of participants	Medium (100-200 attendees)
Website	multimodalagents.github.io

2 Topics covered

Multimodal Agents (MMAs) have shown significant utility across a variety of tasks. The advancements in large-scale foundational models and interactive artificial intelligence have opened up novel capabilities for MMAs, including the ability to predict user actions and devise plans for tasks within specific constraints ([8,9,10,11]). Nevertheless, for MMAs to be genuinely beneficial, they must offer intuitive interaction experiences and adapt to a wide array of environments,

contexts, and modalities. To promote research in this area, we plan to host a workshop on a broad range of topics relevant for multimodal agents including, but not limited to, embodied agents, interactive agents, agents for traditional multimodal tasks, multimodal agent systems and infrastructure, and applications of multimodal agents. We hope that submissions to our workshop will mainly consist of these primary areas, among other relevant topics for the multimodal agent AI community.

3 Program Logistics

We hope to organize our workshop as a hybrid (mixed in-person and virtual attendance) event. To facilitate the virtual experience, we plan to use Stanford's zoom webinar platform for this workshop. Additionally, we will host our website for the workshop at multimodalagents.github.io. As we finalize our list of invited speakers, topics, and accepted papers, we will update our website accordingly.

3.1 Schedule

We plan for our workshop to be a half-day event, and include two invited speaker discussions along with two spotlight presentations for exceptional workshop paper submissions. Afterwards, we will host a poster session for all our accepted works. For a complete outline of our workshop, see Table 2. Due to the size of related workshops, we anticipate approximately 100-200 participants for our workshop, across all speakers, organizers, and workshop attendees.

3.2 Paper Submission

The timeline of the workshop and competition is expected to be as follows:

Table 1. Important Dates for Workshop Authors

Paper submission deadline	July 15, 2024
Notification to authors	August 18, 2024
Camera-ready deadline	August 25, 2024

We will use Microsoft CMT for handling the submissions, and all reviews will be double-blind. The organizing committee will recruit several reviewers and will ensure receiving at least 2 or 3 reviews for each paper. The top 2-3 rated submissions will be selected for spotlight presentation and the rest of the accepted papers will be presented in a poster session. We plan to publish the accepted papers in the ECCV-Workshops Proceedings.

3.3 Awards and Prizes

We plan to have prizes and awards in different categories, which will be determined by the ratings of recruited, human judges. At the moment we have preliminary agreement with Microsoft Research and Salesforce for support of major prize nominations. Due to the large industrial interest in MMA research and our previous experience, we expect no issue in securing funding. However, exact prizes are still to be determined. We list our award and prize categories below:

1. **The best paper award** (2-3 papers, 2 selected for spotlight presentations)
2. **The MMA award** for the agent with the most impressive demo.
3. **The evaluation award** for the human evaluator with the most agent ratings.
4. **The generalist award** for the MMA with the greatest breadth of capability.

Table 2. Our proposed workshop timeline.

Time Slot	Speaker(s)	Details
08:00 - 08:10	Qiuyuan Huang	Opening Remarks
08:10 - 08:40	Jianfeng Gao	Invited Talk
08:40 - 09:10	Katsushi Ikeuchi	Invited Talk
09:10 - 09:25	Coffee Break	
09:25 - 09:35	Award Winner #1	Spotlight #1
09:35 - 09:45	Award Winner #2	Spotlight #2
09:45 - 11:15	Poster session	
11:15 - 12:00	Panel Discussion	Current Limitations and Future Challenges for MMAs
12:00 - 12:10	Juan Carlos Niebles	Closing Remarks and Awards

4 Invited Speakers

For our invited speakers, we have focused on those that have extensive expertise and knowledge in embodied and interactive multimodal agents. In particular, we have already **confirmed** Katsushi Ikeuchi and Jianfeng Gao from Microsoft Research are available to give invited talks. In case of conflict for our invited speakers and to extend our panel discussion to the broader research community, we have **invited** the following speakers to give joint talks with our confirmed speakers or participate in our panel discussion: Chelsea Finn from Stanford University, Dilek Hakkani Tur from the University of Illinois Urbana-Champaign, Demetri Terzopolous from University of California, Los Angeles, Pete Florence from Google Deepmind, Graham Neubig from Carnegie Mellon University, Dorsa Sadigh from Stanford University, and Daniel Fried from Carnegie Mellon University.

5 Related Workshops

The **MMFM Workshop at ICCV 2023** explores large-scale multimodal foundation models that have been pre-trained on massive datasets and can adapt to various tasks with minimal supervision. In contrast, our proposed workshop focuses on multimodal *agent* AI systems. The **Embodied AI Workshop at CVPR 2023** is similar to ours, but with a much greater emphasis on the ability of single agent systems to act within environments. For our workshop, we take a broader view of agent interactions, including areas such as game-playing and interactive ambient intelligence systems. The **AI4ABM Workshop at ICLR 2023** focuses on AI for Agent-Based Modelling. Although AI4ABM also emphasizes the role of AI systems interacting as agents, it mainly considers the more traditional definition of agents within a controlled or simulated environment, and places significantly less emphasis on multimodal agents and the interplay

between multimodal foundation models and generalist agents. The **LLM Agents Workshop at ICLR 2024** studies the usage of LLMs as agentic systems. Although this workshop does explore multimodality as one of their five sub-areas, we believe that MMAs deserve a workshop by their own right, largely due to their broad range of applications beyond natural language and the fact that they need not be tethered to pre-trained LLMs.

6 Diversity Statement

The Multimodal Agents Workshop at ECCV 2024 recognizes the fundamental importance of diversity and inclusion. Therefore, we especially encourage those from groups currently underrepresented in engineering and computer science, including people who self-identify as a woman, African American, Black Hispanic, Latinx, American Indian, Alaska Native, Native Hawaiian, Pacific Islander, person with a disability, and/or LGBTQIA+ to participate in our workshop. To promote diversity in our speakers and panelists, we have invited a diverse set of researchers leading the forefront of Agent-based AI. However, our commitment extends beyond fostering a diverse roster of speakers, panelists, and participants from industry and academia; our organizing committee also possesses a unique blend of academic backgrounds and experiences, ranging from disciplines such as neuroscience, activity recognition, robotics, hospital care, and visual-language understanding. Furthermore, our organizing team consists of people from a broad range of geographical origins, gender identities, organizations, and cultures.

7 Ethical Considerations

Due to the nature of multimodal agents and their ability to act within diverse environments, the type of AI systems developed for our workshop have the potential to cause real, significant impacts at both an individual and societal level. We do not endorse the creation of any AI system that can cause direct harm to humans or society, and strongly believe that AI should be created to serve as a collaborative tool to enhance our own capabilities. We will promptly reject any submission in violation with these principles.

8 Organizing Committee

Our organizing committee includes both senior and junior members from the computer vision community, who are knowledgeable in multimodal agents and their use across diverse domains including robotics, game-playing AI, video understanding, neuroscience, and healthcare. Among the organizing committee, we have authors with extensive experience with multimodal AI systems and their applications as interactive agents. In addition, the organizing team has extensive experience in organizing previous workshops at computer vision conferences [1,2,3,4,5,6,7]. We provide brief backgrounds for each of the organizers below:

- **Zane Durante** Zane is a PhD student in the Stanford Vision and Learning (SVL) lab and Partnership in AI-assisted Care (PAC) advised by Fei-Fei Li and Ehsan Adeli. He is broadly interested in multimodal AI systems, foundation models, and building generalist agents. His PhD work is supported by NSF's Graduate Research Fellowship.

- **Ehsan Adeli** Ehsan is an Assistant Professor at Stanford University. He leads research primarily in the Computational Neuroscience (CNS) Lab and is also affiliated with SVL and PAC. With a Ph.D. in computer science and postgraduate training in biomedical imaging, Ehsan is applying his expertise to solve critical problems in healthcare and neuroscience using datasets with different modalities.

- **Juan Carlos Niebles** Juan Carlos is a Research Director at Salesforce AI Research, a Co-Director of SVL, and an Adjunct Professor of Computer Science at Stanford University. His research interests are in computer vision and machine learning, with a focus on visual recognition of human actions and activities, objects, scenes, and events.

- **Naoki Wake** Naoki Wake is a researcher in the Applied Robotics Research group at Microsoft. His current research involves the development of multimodal perception systems for robots and co-speech gesturing systems. His past research has spanned auditory neuroscience, neuro-rehabilitation, and speech processing. Naoki received his Ph.D. in Information Science and Technology in 2019 from the University of Tokyo.

- **Bidipta Sarkar** Bidipta Sarkar is a senior undergraduate student at Stanford University and a member of Stanford's ILIAD lab. His research focuses on creating AI agents that can interact with their environment and safely work alongside humans and other autonomous agents.

- **Ran Gong** Ran Gong is a PhD student at the UCLA VCLA Lab. His research lies in the intersection of Robotics, Computer Vision, Computer Graphics, and Machine Learning. His research focuses on embodied simulation and interaction with a goal of creating intelligent behaviors that can solve diverse tasks in diverse environments as well as well as the capability of collaborating with humans.

- **Jae Sung Park** Jae Sung is a PhD student advised by Yejin Choi and Ali Farhadi. His research focuses on developing models with multimodal commonsense reasoning. He is interested in equipping models with grounding linguistic concepts to visual modalities, and having them understand multimedia content in a way that humans process the visual information.

- **Yejin Choi** Yejin is a Wisnner-Slivaka Chair and Brett Helsel Professor at University of Washington and Senior Research Manager at Allen Institute of Artifical Intelligence. She has won the Anita Borg Early Career Award in 2018. She was the recipient of MacArthur Fellow foundation fellowship in 2020. She has received outstanding paper award in AAAI 2020, Neurips 2021, ICML 2022, and ACL 2023, and the best paper award in NAACL 2022 and ACL 2023.

- **Fei-Fei Li** Fei-Fei is the inaugural Sequoia Professor in the Computer Science Department at Stanford University, and Co-Director of Stanford's Human-Centered AI Institute. She served as the Director of Stanford's AI Lab from 2013 to 2018. And during her sabbatical from Stanford from January 2017 to September 2018, Dr. Li was Vice President at Google and served as Chief Scientist of AI/ML at Google Cloud. Since then she has served as a Board member or advisor in various public or private companies.

- **Qiuyuan Huang** Qiuyuan Huang is a principal researcher in the deep learning group at Microsoft Research (MSR), Redmond, WA. Her current research interests are mainly in the deep learning, multi-modality, and natural language processing, specifically on Agent AI for Gaming, Robotics and Healthcare; Knowledge-reasoning Intelligence for Interactive AI; Neuro-symbolic Computation for Inference Reasoning; and Large Foundation models for NLP and Multi-modality.

References

1. Activitynet large scale activity recognition challenge, 2016-2022. IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR) Workshop.
2. International challenge on compositional and multimodal perception, 2021-2023. International Conference on Computer Vision (ICCV) and European Conference on Computer Vision (ECCV) Workshop.
3. International workshop on capturing, interpreting & visualizing indoor living spaces (civils), 2023. IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR) Workshop.
4. International workshop on distributed smart cameras, 2021-2022. International Conference on Computer Vision (ICCV) and European Conference on Computer Vision (ECCV) Workshop.
5. International workshop on large scale holistic video understanding (hvu), 2021-2023. IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR) Workshop.
6. Workshop, benchmark, and challenge on human trajectory and pose dynamics forecasting in the wild, 2021. International Conference on Computer Vision (ICCV) Workshop.
7. Workshop on compositionality in computer vision (cicv), 2020. IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR) Workshop.
8. Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Fu, C., Gopalakrishnan, K., Hausman, K., et al.: Do as i can, not as i say: Grounding language in robotic affordances. arXiv preprint arXiv:2204.01691 (2022)
9. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. arXiv preprint arXiv:2307.15818 (2023)
10. Durante, Z., Sarkar, B., Gong, R., Taori, R., Noda, Y., Tang, P., Adeli, E., Lakshmikanth, S.K., Schulman, K., Milstein, A., et al.: An interactive agent foundation model. arXiv preprint arXiv:2402.05929 (2024)
11. Gong, R., Huang, Q., Ma, X., Vo, H., Durante, Z., Noda, Y., Zheng, Z., Zhu, S.C., Terzopoulos, D., Fei-Fei, L., et al.: Mindagent: Emergent gaming interaction. arXiv preprint arXiv:2309.09971 (2023)