Interviews with 2019 AAAI/ACM SIGAI Doctoral Dissertation Awardees

DOI: 10.1145/3795125.3795129

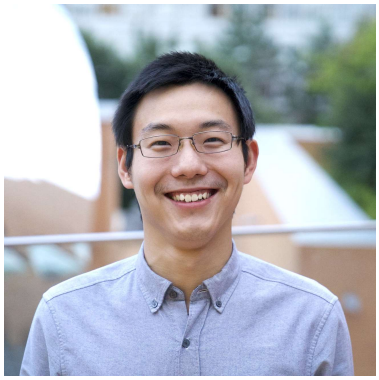
The Joint AAAI/ACM SIGAI Doctoral Dissertation Award recognizes and encourages superior research and writing by doctoral candidates in artificial intelligence. The award is presented annually at the AAAI Conference on Artificial Intelligence, and the winner is invited to present a talk at the conference. Before diving into the winners of the 2025 Award, we thought we'd catch up with some former awardees.

In 2019, Jiajun Wu won the award for his work "Learning to See the Physical World", while Aishwarya Agrawal achieved an honourable mention for her dissertation entitled "Visual Question Answering and Beyond". We interviewed them both to learn about where their research has taken them since then, and what future avenues they are looking to in the field of computer vision.

2019 AAAI/ACM SIGAI Dissertation Award Winner - Jiajun Wu

Massachusetts Institute of Technology, Dissertation Title: "Learning to See the Physical World"

DOI: 10.1145/3795125.3795130



Jiajun Wu is an Assistant Professor of Computer Science and, by courtesy, of Psychology at Stanford University, working on computer vision, machine learning, robotics, and computational cognitive science. He received his PhD in EECS from MIT. Wu's research has been recognized through the Young Investigator Programs (YIP) by ONR and by AFOSR, the NSF CAREER award, the Okawa research grant, the AI's 10 to Watch by IEEE Intelligent Systems, paper awards and finalists at ICCV, CVPR, SIGGRAPH Asia, ICRA, CoRL, and IROS, dissertation awards from ACM, AAAI, and MIT, and the 2020 Samsung AI Researcher of the Year.

Q: What is your research area?

My research topic, at a high level, hasn't changed much since my dissertation. It has always been the problem of physical scene understanding---building machines that see, reason about, and interact with the physical world. Besides learning algorithms, what are the levels of abstraction needed by AI systems in their representations, and where do they come from? I aim to answer these fundamental questions, drawing inspiration from nature, i.e., the physical

world itself, and from human cognition.

Q: What is the context of your work?

Building machines with visual, physical intelligence has been a north star for AI for decades. Despite progress, physical scene understanding remains unsolved, as it requires holistic interpretation of geometry, physics, and functionality---beyond the scope of any single discipline. Data for these domains remain scarce; simply scaling models up is thus infeasible. We need proper representations and learning paradigms that enable data-efficient, flexible, generalizable physical scene understanding.

Q: What is your methodology?

My approach to constructing representations of the physical world is to integrate bottom-up recognition models and efficient inference algorithms with top-down graphical models, generative models, and neural, analytical (often differentiable), and hybrid simulation engines. My research develops these techniques (e.g., proposing new deep networks and hybrid physical simulators); we also further explore innovative ways to combine them, building on cross-disciplinary studies.

Q: How is your research developing now?

In our research, we always aim to infer, represent, and use physical world structure from raw visual data, without compromising the expressiveness of neural networks. Recently, with the rapid development of visual AI models, we have continued to investigate what role such structural information plays, or whether we still need it at all. Our recent efforts in this direction can be categorized into two technical paths: leveraging physical world structure as powerful inductive biases, or grounding pre-trained vision or multi-modal foundation models onto the physical world. We can now build visual intelligence that infers object shape, texture, material, and physics, as well as scene context, with applications in controllable, action-conditioned 4D visual world reconstruction, generation, and interaction.

Q: What are the main applications of your research?

The main use of computer vision is of course for robotics, but we can also use it for entertainment (movies, games), design, and creativity. For example, one of our recent papers attracted a lot of interest from game designers: https://kyleleey.github.io/WonderPlay/

Q: How have you seen your field evolve in recent years?

AI advances, or “hype,” have sparked many discussions about the “identity crisis” of academia—industry jobs have become much more attractive to fresh PhDs; some people are questioning the role of academic research given the extreme imbalance in resources across many dimensions. We as academic researchers have to rethink the value, focus areas, and perspectives of academic research that are still worth exploring (which I believe remain many) so that fundamental, long-term research will continue to shine.

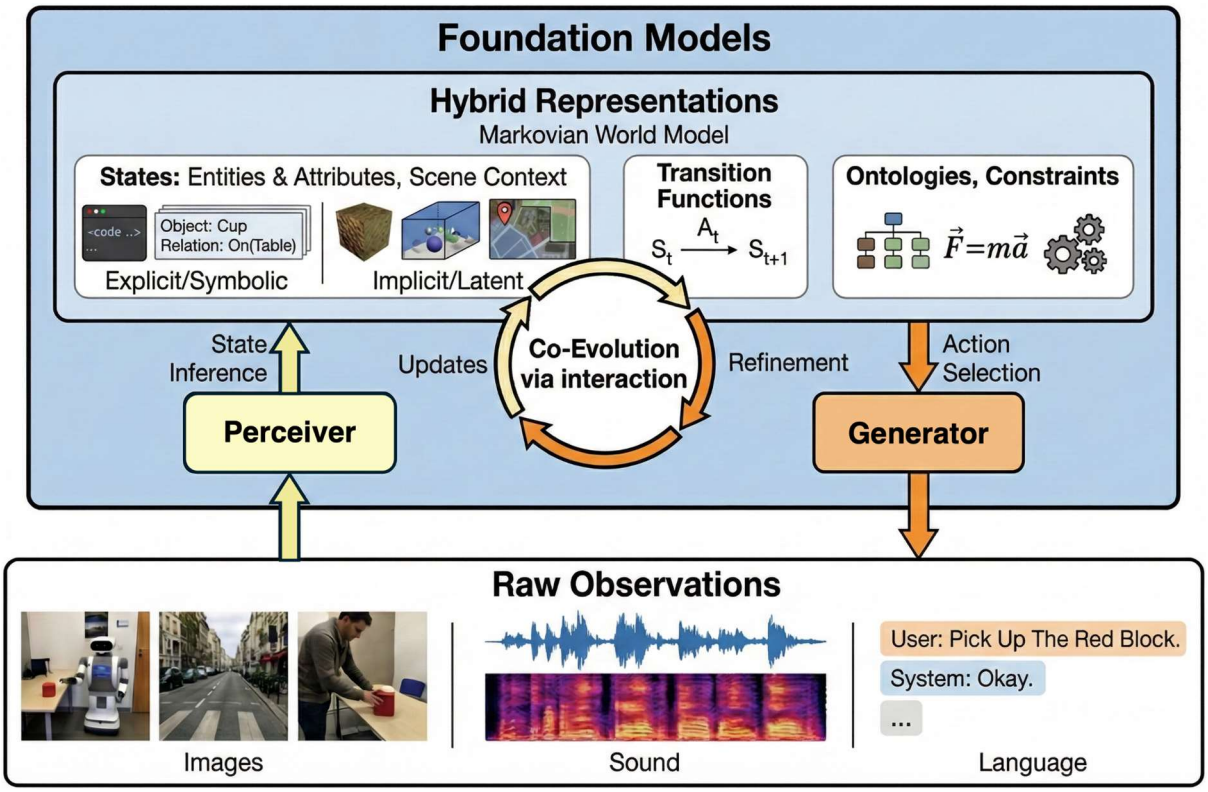


Image Credits: *Jiajun Wu, Yunzhi Zhang, Hong-Xing Yu, Joy Hsu, Jiayuan Mao. Discovering Hybrid World Representations with Co-Evolving Foundation Models. In Proceedings of the Annual AAAI Conference on Artificial Intelligence, Emerging Trends in AI (ETA) Track, 2026.*

Q: Which future directions or open questions excite you the most?

In light of the exciting advances in foundation models, we have been exploring how to adapt them for physical world modeling. Beyond the two paradigms I mentioned above, we may benefit from continual learning to refine the discovered physical worlds and, possibly, the foundation models themselves through interactions with the real world. The continual learning paradigm, including an iterative cycle of perception, interaction, and symbolic abstraction, will better leverage the commonsense knowledge from foundation models.

Such interactive learning is mutually beneficial. The discovered physical world model continues to improve through the interpretation of interaction results by foundation models; at the same time, new knowledge gained from interactions is fed back into foundation models, enabling their continued pre-training for better compression, summarization, and future reasoning. This establishes a co-evolving loop in which both world models and foundation models become more capable.