Building Physically Plausible World Models

Homanga Bharadhwaj[1], **Boyuan Chen**[2], **Yilun Du**[34], **Hiroki Furuta**[45], **Ruiqi Gao**[4], **Hamidreza Kasaei**[6],

Sean Kirmani[4], **Kuang-Huei Lee**[4], **Ruoshi Liu**[7], **Zeyi Liu**[8], **Fei-Fei Li**[89], **Carl Vondrick**[7], **Wenhao Yu**[4]

[1]CMU, [2]MIT, [3]Harvard, [4]Google DeepMind, [5]UTokyo, [6]University of Groningen, [7]Columbia, [8]Stanford, [9]World Labs

1 Introduction

The goal of this workshop is to exchange ideas and establish communications among researchers working on building generalizable world models that describe how the physical world evolves in response to interacting agents (e.g. human and robots). Large-scale datasets of videos, images, and text hold the key for learning generalizable world models that are visually plausible. However, distilling useful physical information from such diverse unstructured data is challenging and requires careful attention to data curation, developing scalable algorithms, and implementing suitable training curricula. On the other hand, physics-based priors can enable learning plausible scene dynamics but it is difficult to scale to complex phenomenon that lack efficient solvers or even governing dynamic equations. Developing general world models that can simulate complex real-world phenomenon in a physically-plausible fashion can unlock enormous opportunities in generative modeling and robotics, and would be of wide interest to the larger AI community, and we believe this workshop falls at an ideal timing given recent significant progress in both video-modeling models and physics-based simulation. This workshop aims to bring together researchers in machine learning, robotics, physics-based simulation, and computer vision broadly aspiring to build scalable world models by utilizing internet data, simulation, and beyond in myriad ways.

1.1 Topics of interest for the workshop

Our workshop will focus on topics including but not limited to the following:

- **Controllable Video Generation and Generative Simulations.** How can we improve fine-grained control in video generation and integrate it with world models conditioned on low-level actions [1, 2, 3]?
- **Incorporating physics priors.** How can we leverage physics prior to empower learned world modesl with physical realism?
- **Dynamic 3D Reconstruction.** How can we generalize 3D reconstruction using web data while preserving scene consistency and controlled motion of dynamic elements [4, 5, 6, 7, 8]?
- **Applications to Robotics and Time-Series Predictions.** How can we use generic dataset such as web video and text to build shared world models that can synthesize physically-plausible results for applications such as robotics [9, 10, 11] [12, 13]?
- **Special Considerations: Data Curation, Hallucination, and Broader Implications.** How do dataset biases impact learned world models, and how can we mitigate them [14, 15, 16, 17, 18]?

1.2 Relation to Previous Workshops

Our workshop is related to a few prior workshops in machine learning conferences like 1) Generative Modeling and Model-Based Reasoning for Robotics and AI (ICML 2019), 2) Foundation Models for Decision Making (NeurIPS 2022, 2023), 3) Generative Models for Decision Making (ICLR 2024), 4) Workshop on Foundation Models in the Wild (ICML 2024). Unlike 1) we are interested in a broader interpretation of world models beyond model-based reasoning (for example to include conditioned video/3D models), and unlike 2), 3), 4) we have a more focused scope of investigating only world models (as opposed to other generative models and foundation models) and their applications to robotics and physics-based reasoning tasks.

2 Invited Speakers and Logistics

Our list of speakers and panelists consists of researchers across different geographical regions, from both industry and academia, and with diversity in research topics, gender, ethnicity, and seniority.

- **Sherry Yang** (New York University, she/her) **confirmed** speaker and panelist (Scalable Decision Making).
- **Tim Brooks** (Google DeepMind, he/him) **confirmed** speaker and panelist (Video Generation).
- **Agrim Gupta** (Google DeepMind, he/him) **confirmed** speaker and panelist (Vision-Language Models).

ICML2025 Workshop, Vancouver.

- **Shuran Song** (Stanford University, she/her) **confirmed** speaker and panelist (Robot Learning).
- **Hao Su** (University of California San Diego, he/him) **confirmed** speaker and panelist (Simulation).
- **Justin Johnson** (World Labs, he/him) *tentative* speaker/panelist (3D Vision).
- **Beom Joon Kim** (KAIST, he/him) *tentative* panelist (Mobile Manipulation).

Panel Discussions. Further, we will have two panel discussions at least one of which will be framed as a debate on a topic related to the workshop theme. This will involve a subset of the speakers above and some other invited panelists (**2-3 additional invites will be sent out a bit closer to the workshop date**). We will invite the lead authors of the two Best Paper Award winners to also be a panel member for one of the sessions.

Contributed (Spotlight) Talks. In addition to the invited talks and panel discussions, we will have 6 spotlight talks (best paper award finalists from each of the two categories) selected from the paper submissions to the workshop. These will be expected to be delivered by the lead authors of the respective papers and provide a platform to showcase new work led by junior researchers.

2.1 Expected Attendance and Format

Our workshop will showcase high-quality invited talks from established figures and rising talents across academia, industrial labs, and research institutions. We are committed to inviting experts from a variety of backgrounds (e.g., academia, industry), ethnicities, and career stages (e.g., early-career researchers, senior faculty) to ensure a well-rounded and diverse perspective on the field. We anticipate a moderate workshop size, with around 100-300 attendees. While we highly encourage in-person participation, the workshop will also support virtual attendance for those unable to attend in person (e.g., due to visa issues, travel constraints, and any unplanned events). The tentative schedule and timeline are as follows.

- 08:50 - 9:00 **Welcome/Opening Remarks**
- 09:00 - 09:30 **Invited Talk 1**: (including 5 min Q&A)
- 09:30 - 10:00 **Invited Talk 2**: (including 5 min Q&A)
- 10:00 - 10:30 **Contributed Talks**: 10 min. presentations each (one presentation by a best paper)
- 10:30 - 11:30 **Poster Session with Coffee:** We will encourage folks to make new connections and chat!
- 11:30-12:15 **panel discussion / debate 1**
- 12:15 - 13:30 **lunch break**
- 13:30 - 14:00 **Invited Talk 3**: (including 5 min Q&A)
- 14:00 - 14:30 **Invited Talk 4**: (including 5 min Q&A)
- 14:30 - 15:00 **Contributed Talks**: 10 min. presentations each (one presentation by a best paper)
- 15:00 - 16:00 **Poster Session with Coffee:** We will encourage folks to make new connections and chat!
- 16:00 - 16:30 **Invited Talk 5**: (including 5 min Q&A)
- 16:30 - 17:00 **Invited Talk 6**: (including 5 min Q&A)
- 17:00 - 17:45 **panel discussion / debate 2**

2.2 Paper Submission

We ensure inclusivity by accepting two types of submissions:

- **Short Papers / Extended Abstracts (max 3 pages)** that showcase an interesting idea, preliminary results (theoretical or empirical), an interesting application in the context of robotics, or an original idea that did not pan out in practice. The aim of this submission track is to lower the barrier of entry and ensure that junior researchers and researchers with limited resources are given the opportunity to share their work with the ICML community. The top three short papers will be invited for a spotlight talk and the remaining accepted papers will be presented in a poster session.
- **Full submissions (max 8 pages)** that contribute original work in line with the overall theme of the workshop. There will be three award candidates each with a spotlight talk, and the rest of the papers will be presented in a poster session.

The workshop papers are non-archival, and we would be okay accepting submissions that have already been submitted to or accepted by other venues. We will host a poster session featuring around 20-30 posters and are tentatively planning to host it in the workshop room. For reviewing papers to be accepted by the workshop we will recruit external reviewers and form a program committee, and the entire review process will be double-blind.

References

[1] W. Yan, Y. Zhang, P. Abbeel, and A. Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

[2] D. Kondratyuk, L. Yu, X. Gu, J. Lezama, J. Huang, R. Hornung, H. Adam, H. Akbari, Y. Alon, V. Birodkar, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023.

[3] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[4] I. Lee, B. Kim, and H. Joo. Guess the unseen: Dynamic 3d scene reconstruction from partial 2d glimpses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1062–1071, 2024.

[5] Q. Wang, V. Ye, H. Gao, J. Austin, Z. Li, and A. Kanazawa. Shape of motion: 4d reconstruction from a single video. *arXiv preprint arXiv:2407.13764*, 2024.

[6] B. Roessle, N. Müller, L. Porzi, S. R. Bulò, P. Kontschieder, A. Dai, and M. Nießner. L3dg: Latent 3d gaussian diffusion. *arXiv preprint arXiv:2410.13530*, 2024.

[7] K. Gao, Y. Gao, H. He, D. Lu, L. Xu, and J. Li. Nerf: Neural radiance field in 3d vision, a comprehensive review. *arXiv preprint arXiv:2210.00379*, 2022.

[8] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[9] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[10] I. Radosavovic, T. Xiao, B. Zhang, T. Darrell, J. Malik, and K. Sreenath. Real-world humanoid locomotion with reinforcement learning. *Science Robotics*, 9(89):eadi9579, 2024.

[11] T. Miki, J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter. Learning robust perceptive locomotion for quadrupedal robots in the wild. *Science robotics*, 7(62):eabk2822, 2022.

[12] M. Deisenroth and C. E. Rasmussen. Pilco: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on machine learning (ICML-11)*, pages 465–472, 2011.

[13] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

[14] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.

[15] Z. Fan, M. Parelli, M. E. Kadoglou, X. Chen, M. Kocabas, M. J. Black, and O. Hilliges. Hold: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 494–504, 2024.

[16] J. L. Schonberger and J.-M. Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.

[17] N. Karaev, I. Rocco, B. Graham, N. Neverova, A. Vedaldi, and C. Rupprecht. Cotracker: It is better to track together. *arXiv preprint arXiv:2307.07635*, 2023.

[18] R. Baeza-Yates. Bias on the web. *Communications of the ACM*, 61(6):54–61, 2018.

A Organizer Biographies

We assemble a diverse team of experts from robotics, computer vision, and reinforcement learning (RL) across two continents within both industry and academia. The team comprises of senior and junior researchers and is diverse in terms of ethnicity, gender, seniority, institution, and research expertise. They have been at the forefront of developing scalable algorithms for RL, developing robot learning algorithms utilizing web video datasets, and contributing to fundamental advances in 3D reconstruction and generation. Many members of the team have significant experience organizing prior workshops in top machine learning, computer vision, and robotics conferences (e.g. ICML, ICLR, ECCV, RSS etc. [1]) and the diverse and comprehensive experiences makes the team a strong candidate for organizing the proposed inter-disciplinary workshop.

- **Homanga Bharadhwaj** is a final-year PhD student in Carnegie Mellon University. His research goal is to develop embodied AI systems capable of helping us in the humdrum of everyday activities within messy rooms, offices, and kitchens, in a reliable, compliant, and scalable manner without requiring significant robot-specific data collection and task-specific heuristics. Homanga was named a Future Leader in Robotics and AI by the University of Maryland in 2025 and was selected as a Meta AI Mentorship (AIM) Fellow in 2022. His research has received a Best Paper in Robot Manipulation Finalist at ICRA '24, the Best Conference Paper Award at ICRA '24, the Outstanding Presentation award at NeurIPS '23 Robot Learning Workshop, and has been covered by several media outlets like TechCrunch, IEEE Spectrum, and VentureBeat among others. [Google Scholars, website]
- **Boyuan Chen** is a fourth-year PhD student at MIT working with Prof Vincent Sitzmann and Prof Russ Tedrake. He is interested in model-based reinforcement learning, generative world model and their applications in embodied intelligence. Boyuan hopes to leverage video world models trained on internet-scale data as planners for general-purpose robots, replicating LLM's success but for the visual world. Previously, Boyuan also interned at Google Deepmind and Google X, working on equipping Google's foundation models with spatial reasoning capabilities. [Google Scholars, website]
- **Yilun Du** is an assistant professor in computer science at Harvard University and a Senior Research Scientist at Google Deepmind. His research focuses on developing intelligent embodied agents. In the past, he has organized a number of workshops including Foundation Models for Decision (NeurIPS 2022 / 2023), Generative Models in Decision Making (ICLR 2024), Compositional Learning (NeurIPS 2024), Safe Generative AI (NeurIPS 2024), and Geometric and Algebraic Structures for Robotics (RSS 2024). [Google Scholars, website]
- **Hiroki Furuta** is a Research Scientist at Google DeepMind Japan and a final-year PhD student in The University of Tokyo. His research interests are focused on developing agent and alignment in generative AI through deep reinforcement learning; covering the topic from world models for game and robotics to reinforcement learning from AI feedback for video diffusion models. He was recognized as a Forbes JAPAN 30 UNDER 30 2023 honoree for his outstanding research achievements in Japan. Previously, he has organized Ecological Theory of RL Workshop at NeurIPS 2021. [Google Scholars, website]
- **Ruiqi Gao** is a Senior Research Scientist at Google Deepmind. Her research interests are in statistical modeling and learning, with a focus on scalable training and inference algorithms of deep generative models such as diffusion models. She has organized a number of workshops and tutorials in the past including Structured Probabilistic Inference and Generative Modeling (ICML 2024), Efficient and On-Device Generation (CVPR 2024) and Latent Diffusion Models (NeurIPS 2023). [Google Scholars, website]
- **Hamidreza Kasaei** (he/him) is an Assistant Professor in the Department of Artificial Intelligence at the University of Groningen, the Netherlands. He runs the IRL-Lab focusing on Lifelong Interactive Robot Learning in 3D Object Perception, Grasp Affordance, and Object Manipulation. He holds the Google Research Award 2023, Gratama Science Award 2022, and the Robotics: Science and System Pioneers Award 2019. Hamidreza has organized workshops at IROS'19, RSS'19, '22, NeurIPS'21, '22, '23, ICLR'25, and has been an associate editor for the RA-L journal, ICRA, and IROS since 2020. He was named as an outstanding associate editor for the RA-L in 2023. Google Scholars, website]
- **Sean Kirmani** is a Senior Research Scientist at Google DeepMind working on problems in vision, language, and action. His research interests include improving 3D vision in foundation models and simulating robots at scale with world models. He also was the technical lead for semantic perception at The Everyday Robot Project within Google[x]. He holds degrees in electrical engineering and computer science from The University of Texas at Austin where he worked in several robotics labs. [Google Scholars, website]
- **Kuang-Huei Lee** is a Staff Research Scientist at Google DeepMind. His research interests are focused on the creation of general cognitive agents in the physical and virtual worlds. His recent research spans deep generative models, reasoning, reinforcement learning, robotics, and AI for chip design. Prior to joining Google in 2019, Kuang-Huei spent 3 years at Microsoft. He received his graduate degree in Computer Science from Carnegie Mellon University,

[1] https://3d-in-the-wild.github.io/ https://tarl2019.github.io/ https://social-intelligence-human-ai.github.io/

and his undergraduate degree in Mechanical Engineering from National Taiwan University. His work has been widely published at NeurIPS, ICML, ICLR, CoRL, RSS, IROS, CVPR, ECCV, EMNLP, etc., with recognition of a CoRL Special Innovation Award and an IROS Best Paper Finalist. [Google Scholars, website]

- **Ruoshi Liu** is a Ph.D. student in computer science at Columbia University. His research focuses on developing embodied intelligent systems, with interests in 3D/4D perception, generative models, and robot learning. His work received Best Paper Finalist at ECCV 2024 and news coverage from New Scientists, Communications of the ACM, Hacker News, VentureBeat, and TechCrunch etc. Open-source models and datasets he developed have been downloaded and used more than a million times by other researchers and engineers in the field. He has organized the 4D Vision workshop at CVPR 2025 and NYC Computer Vision Day in the past 3 years. [Google Scholars, website]
- **Zeyi Liu** is a third-year PhD student at Stanford University, USA. Her research focuses on robot perception and manipulation. She's interested in enabling embodied agents to better perceive and understand the complex and diverse physical world through multimodal data (e.g., vision, language, audio, force), which facilitates robust and generalizable policy learning. To this end, she works on data collection methods and policy learning frameworks that efficiently learn from multimodal inputs, allowing robots to acquire more precise and robust manipulation skills. [Google Scholars, website]
- **Fei-Fei Li** is the inaugural Sequoia Professor in the Computer Science Department at Stanford University, and a Founding 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. Li was named in the Time 100 AI Most Influential People list in 2023 and received the Intel Lifetime Achievements Innovation Award in the same year for her contributions to artificial intelligence. She was elected as a member of the National Academy of Engineering[18] and the National Academy of Medicine in 2020, and the American Academy of Arts and Sciences in 2021. [Google Scholars, website]
- **Carl Vondrick** is a Associate Professor in the computer science department at Columbia, and his research studies computer vision, machine learning, and their applications. He was previously a research scientist on the machine perception team at Google, and a visiting researcher at Cruise. He completed my PhD at MIT in 2017 advised by Antonio Torralba and his BS at UC Irvine in 2011, where he got his start working with Deva Ramanan. He received the 2024 PAMI Young Researcher Award and the 2021 NSF CAREER Award. He is also the Senior Program Chair for ICLR 2025. [Google Scholars, website]
- **Wenhao Yu** is a Staff Resesarch Scientist at Google DeepMind. His research lies at the intersection between robotics, machine learning, and computer animation. His work combines reinforcement learning, control, foundation model, and simulation, to create intelligent mobile robots that can perform real-world useful tasks in human-centered environments. Wenhao obtained his PhD in Computer Science from Georgia Institute of Technology, USA. [Google Scholars, website]

Contact Organizers: Wenhao Yu, Kuang-Huei Lee ({magicmelon, leekh}@google.com), Homanga Bharadhwaj (hbharadh@cs.cmu.edu)

B Diversity Statement

We actively ensure the diversity of the workshop participants.

1. The workshop is proposed by a diverse organizing committee, in the following aspects.
   - **Gender**: *female* (Ruiqi, Zeyi, Fei-Fei) and *male* (Homanga, Boyuan, Yilun, Hamidreza, Sean, Kuang-Huei, Ruoshi, Carl, Wenhao, Hiroki)
   - **Seniority**: *Professor* (Fei-Fei), *Associate Professor* (Carl), *Assistant Professor* (Yilun, Hamidreza), *Research Scientist with more than 5 years of experience* (Sean, Kuang-Huei), *Research Scientist with 3-5 years of experience* (Ruiqi, Wenhao), *Research Scientist with less than 3 years of experience* (Hiroki) and *PhD students* (Boyuan, Zeyi, Homanga, Ruoshi).
   - **Institutional affiliation**: *Academia* (Fei-Fei, Carl, Boyuan, Yilun, Hamidreza, Zeyi, Homanga, Ruoshi, Hiroki) and *Industry* (Fei-Fei, Yilun, Sean, Kuang-Huei, Ruiqi, Wenhao, Hiroki)
   - **Geographical location**: *Europe* (Hamidreza), *Asia* (Hiroki), *North America West Coast* (Ruiqi, Zeyi, Fei-Fei, Sean, Kuang-Huei, Wenhao), *North America East Coast* (Homanga, Boyuan, Yilun, Ruoshi, Carl)
   - **Research area**: The organizing committee covers a diverse set of areas, including generative modeling, video generation, computer vision, simulation, reinforcement learning, multimodal large langauge models, and robotics.
2. The speaker and panelist lineup presents diversity in the following aspects:
   - **Gender**: *female* (Sherry Yang, Shuran Song) and *male* (Tim Brooks, Agrim Gupta, Hao Su, Justin Johnson, Beom Joon Kim)

- **Seniority**: *Professor* (Hao Su), *Associate Professor* (Beom Joon Kim), *Assistant Professor* (Shuran Song, Justin Johnson, Sherry Yang), *Research Scientist more than 5 years of experience* (Sherry Yang), *Research Scientist with 3-5 years of experience* (Tim Brooks) and *Research Scientist with less than 3 years of experience* (Agrim Gupta).
- **Institutional affiliation**: *Academia* (Hao Su, Beom Joon Kim, Shuran Song, Sherry Yang) and *Industry* (Tim Brooks, Agrim Gupta, Justin Johnson)
- **Geographical location**: *Asia* (Beom Joon Kim), *North America West Coast* (Hao Wu, Shuran Song, Tim Brooks, Agrim Gupta, Justin Johnson), *North America East Coast* (Sherry Yang)
- **Research area**: The research area of the invited speakers span a wide spectrum, including video generation, vision-language modeling, 3D Vision, computer simulation, and robotics.

3. We will also issue a broad call for posters and actively promote workshop participation to reach a wide and diverse audience. In particular, we will engage with institutions and organizations serving underrepresented groups—such as Black in AI and Women in Machine Learning—to ensure inclusive outreach and encourage broader involvement in the workshop.
4. We will recruit program committee members from both academia and industry across all continents with research expertise spanning generative modeling, video generation, computer vision, simulation, reinforcement learning, robotics, and robot learning.