NVIDIA AI Podcast:

January 7, 2025: NVIDIA's Ming-Yu Liu on How World Foundation Models Will Advance Physical AI

Featuring Ming-Yu Liu, Vice President of Research at NVIDIA

Transcript:

Noah Kravitz: Hello and welcome to the NVIDIA AI Podcast. I'm your host, Noah Kravitz. NVIDIA CEO Jensen Huang recently keynoted the CES Consumer Electronics show conference in Las Vegas, Nevada. Amongst the many exciting announcements Jensen talked about was NVIDIA Cosmos. Cosmos is a development platform for world foundation models, which I think we're all going to be talking a lot about in the coming months and years. What is a world foundation model? Well, thankfully we've got an expert here to tell us all about it. Ming-Yu Liu is Vice President of Research at NVIDIA. He's also an IEEE Fellow and he's here to tell us all about world foundation models, how they work, what they mean, and why we should care about them going forward. So without further ado, Ming-Yu, thank you so much for joining the NVIDIA AI Podcast and welcome.

Ming-Yu Liu: It's great to be here.

Kravitz: So let's start with the basics if you would. What is a world foundation model?

Liu: Sure. So, world foundation models are deep learning-based space-time, visual simulator that can help us look into the future. You can simulate visits, it can simulate people's intentions and activities. It's like data strain of AI. Imagine many different environments and can simulate the future so we can make good decisions based on this simulation. We can leverage world foundation models, imagination, and simulation capability to help train physical AI agents. We can also leverage this capability to help the agent make good decision during the inference time. You can generate a virtual world based on test prompts, image prompts, video prompts, action points, and the layer combinations. So we call it a world foundation model because it can generate many different worlds and also because it can be customized through different physical AI setups to become a customized world model, right? So different physical AI have different number of cameras and different locations. So we want the world foundation model to be customizable for

different physical AI setups so they can use in their settings.

Kravitz: So I want to ask you kind of how a world model is similar or different to an LLM and other types of models. But I think first I want to back up a step and ask you, how is a world model similar or different to a model that generates video? Because my understanding, and please correct me where I'm wrong, my understanding is that you can prompt a world model to generate a video, but that video is generated based on the things you were talking about, based on understanding of, you know, physics and other things in the physical world and it's a different process. So I don't know what the best way is to kind of unpack it for the listeners, but one place to start might be how does a world model differentiate from an LLM or a generative AI video model?

Liu: So a world model is different to LLM in the sense that LLM is focused on generating text description.

Kravitz: Right.

Liu: It generates understanding, right? A world model is generating simulation. And the most common form of simulation is videos.

Kravitz: Okay.

Liu: So they are generating pixels and so world models and video foundation models, they are related. And video foundation model is a general model that generates videos. It can be for creative use cases, it can be for other use cases. In world models we are focusing on this aspect of video generation. Based on your current observation and the intention of the actors in your world, you roll out the future.

Kravitz: Right.

Liu: Yeah. So they are related, but with a different focus.

Kravitz: Gotcha. Thank you. So why do we need the world models? I mean, I think I know part of the answer to the question. We're talking about simulating physical AI and all of these amazing things. But what's the, you know, tell us about the need for

world foundation models from your perspective.

Liu: So I think world foundation models is important to physical AI developers. You know, physical AI are systems with AI deployed in the real world, right? And different to digital AI, these physical AI systems that interact with the environment can create damages, right? So this could be real harm, right?

Kravitz: Right. Right. So a physical AI system might be controlling a robotic arm or some other piece of equipment, changing the physical world.

Liu: Yeah. I think there are three major use cases for physical AI.

Kravitz: Okay.

Liu: It's all around simulation. The first one is when you train a physical AI system, you train a deep learning model. You have a thousand checkpoints. Do you know which one you want to deploy? Right?

Kravitz: Right.

Liu: And if you deploy individually, going to be very time-consuming and it's bad, it's going to damage your kitchen, right? So with a world model, you can do verification in the simulation.

Kravitz: Right.

Liu: So you can quickly test out these policies in many, many different kitchens. And before you know, you deploy the real kitchen. And after this verification step, you may be narrow down to three checkpoints and then you do the real deployment so you can have an easier life with your physical AI.

Kravitz: It reminds me of when we've had podcasts about drug discovery and the guests talking about the ability to simulate experiments and different molecular combinations and all of that work so that they can narrow it down to the ones that are worth trying in the actual the physical lab, right? So it sounds like a similar, like just being able to simulate everything and narrow it down must be such a huge

advantage to developers.

Liu: Yeah. And second of application is world model, if you can predict the futures, you have some kind of understanding of physics, you might know the action required to drive the world toward that future. And the policy model, the typical one deploying physical AI is all about predicting the action. The right action given the observation. So world model can be used as initialization to the parsing model. And then you know, you can train the parsing model with less amount of data because the world model is already pretrained with many different observations that span the data sets.

Kravitz: So without a world model, what's the procedure of training a policy like?

Liu: So one procedure is you collect data and then you start to do the supervised fine-tuning.

Kravitz: Right.

Liu: And then you may use. Yeah.

Kravitz: So it's hands on, it's manual, you have to get all the data. It's a lot.

Liu: Yeah, yeah. And third one is when world model is good enough, highly accurate and fast. You know, before robot taking any actions, you just simulate different features. And check which you want really achieve your goal and take that one. I have a data strain next to you before you making any decision, wouldn't it be great?

Kravitz: You mentioned accuracy when the models are fast enough and accurate enough. And I don't know if it's a fair question to ask, so ask it, interpret it the best way. But how do you determine accuracy or measure accuracy on a world model? And is there a benchmark that you know, different benchmarks you need to hit to deploy in different situations or how does that work?

Liu: Yeah, it's a great question. So I think a world model development is still in its infancy.

Kravitz: Right, of course.

Liu: So people are still trying to figure out the right way to measure the world model performance. And I think there are several aspects a world model must have. One is follow the law of physics. When you're dropping a ball, you should predict it in the right position based on the physics laws.

Kravitz: Right.

Liu: And also in the 3D environment, we have to have object permanence. So when you turn back and come back, the object should remain there, right? Without any other players, it should remain in the same location. So there are many different aspects I think we need to capture. And I think an important part for the research community is to come out with the right benchmark so that the community can move forward in the right location to democratize this important area.

Kravitz: Right. So speaking of moving forward, maybe we can talk a little bit or you can talk a little bit about Cosmos and what was announced at CES.

Liu: So in CES, Jensen announced the Cosmos world model development platform. It's a developer-first world model platform. So in this platform there are several components. One is pretrained world foundation models.

Kravitz: Right.

Liu: We have two kind of world foundation model. One is based on diffusion, the other is based on autoregressive. And we also have tokenizers for the world foundation models. Tokenizers compress videos into tokens so that transformer can consume for their task.

Kravitz: Right, right.

Liu: In addition to these two, we also provide post-training scripts to help physical AI builder to fine-tune the pretrained model to layer physical AI setup. You know some cars have eight cameras and we would like our world foundation model to predict eight views. And lastly, we also have this video curation toolkit. So processing videos allow video is they are ready to computing task. There are many

pieces need to be processed media gather libraries as they're ready to GPU computation code together want to help the world model developers leverage the library curate data. Either they want to build their own world models or fine-tune one based on our pretrained world foundation models.

Kravitz: So the models provided as part of Cosmos, those are open to developers to use. Are they open to other businesses, enterprises?

Liu: Yes. So this is a open-weight development platform. So meaning that the model is open-weight. The model weights are released before commercial use. We feel this is important to physical AI builders, right? So physical AI builders, they need to solve tons of problems to build really useful robots, self-driving cars for our society. There are so many problems and world model is one of them. And those companies, they may not have the resources or expertise to build a world model.

Kravitz: Right.

Liu: NVIDIA care about our developers and we know many of them are trying to make a huge impact in physical AI. So we want to help them. That's why we create this world model development platform for them to leverage so that they can handle other problems and we can contribute our part to the transformation of our society.

Kravitz: Absolutely. I wanted to ask you, can you explain a little bit about the difference between diffusion models and autoregressive models, particularly in this context, why offer both? What are the use cases? And pros and cons.

Liu: So autoregressive model or AR model is a model that predicts token once at a time.

Kravitz: Right.

Liu: Condition on what has been observed, right? So GPT is probably the most popular autoregressive model. We did token at the time. Diffusion, on the other hand, is a model that we did a set of token together and through iteratively remove noises from these initial tokens.

Kravitz: Right, right, right. Okay.

Liu: Yeah. And the difference is that for AR model, with a significant amount of investment in GPT, there are so many optimizations, so they can run very fast.

Kravitz: Right, right, right.

Liu: And deep fusion, because tokens are generated together, so it's easier to have coherent tokens.

Kravitz: Right.

Liu: The generation quality tend to be better. And both of them are useful for physical AI builders. So some of them need speed, some of them need high accuracy. So both are good.

Kravitz: Excellent.

Liu: So far the most successful autoregressive model is based on discrete token prediction, like in GPT. So you pretty much as a set of integers, tokens and you predict them during training. And in the case of world foundation models, it means you have to organize videos into a set of integers. And you can imagine it's a challenging compression task. And because of this compression, the autoregressive model tend to struggle more on the accuracy. But it has other benefits. For example, its setting is easier integrated into the physical AI setup.

Kravitz: Got it. I'm speaking with Ming-Yu Liu. Ming-Yu is vice president of research at NVIDIA and he's been telling us about world foundation models, including the announcement of NVIDIA Cosmos, the developer platform for world models that was announced during Jensen's CES keynote. So we've been talking a lot about. You've been explaining what a world model is, how it's similar and different to other types of AI models. Just now, the difference between autoregression and diffusion. Let's kind of change gears a little bit and talk about the applications. How will Cosmos, how are our world foundation models going to impact industries?

Liu: Yeah. So we believe that first of all, the world foundation model can be used as a synthetic data generation engine to generate different synthetic data. And like

what I said earlier, the world model can also be used as a policy evaluation tool to determine which checkpoint or which policy is a better candidate for you to test out in the physical world. And also if you can predict the future, it probably can reconfigure to predict the action toward that future. So as a policy pretraining initialization.

Kravitz: Right, right.

Liu: And also to have a data strain next to you before any endeavor. So during the next time future rollout and pick the best decision for each moment.

Kravitz: Are there particular industries I know work in factories and industrial work, anything involving robotics, but are there specific industries that you see benefiting from world models? Maybe sooner than others?

Liu: Yes. I think the self-driving car industry and the human robot industry will benefit a lot from these world model developments. They can simulate different environment that will be difficult to have in the real world to make sure the agent is behave expectedly.

Kravitz: Right.

Liu: So I think these are two very exciting industries the world models can impact.

Kravitz: And NVIDIA obviously has a long history, as you were saying, of, you know, it's not just about rolling out the hardware. There's the software, the stack, the ecosystem, all the work to support developers, because if the devs aren't building, you know, world-changing things with the products, then there's a problem, right? What are some of the partnerships, the ecosystems relative to world foundation models? And maybe there are some partners who are already doing some interesting stuff with the tech you can talk about.

Liu: Yes, we are working with a couple of humanoid companies and self-driving car companies including 1X, Huobi, vAuto, XPENG, and many others.

Kravitz: Right.

Liu: So NVIDIA believe in suffering. We believe that true greatness come from suffering. So working with our partners, we can look at the challenges they are facing to experience their pain and to help us to build a world model platform that is really beneficial to them.

Kravitz: Fantastic. Yeah.

Liu: So I think this is important part to make the field move faster.

Kravitz: Absolutely. All right, so you talked about being able to predict the future and you talked about just now that things moving faster. What do you see on the horizon? What's next for world foundation models? Where do you see this going in the next you know, five years or adjust that timeframe to whatever makes sense.

Liu: So I'm trying to build a world model now, try to predict the future.

Kravitz: Exactly. Yep. Putting you on the spot.

Liu: Yes. I believe we are still in the infancy of world foundation model development. The model can do physics to some extent, but not well or robust enough. That's the critical point to make a huge transformation. It's useful, but we need to make it more useful. So the field of AI advance very fast. So from GPT3 to ChatGPT is just a year or two.

Kravitz: Right. Yeah, we forget it's all going so quickly.

Liu: Yeah, it's going so fast. I believe physical AI development will be very fast too because the infrastructure for large-scale model has been established through this large language model transformation, right? And there's a strong need to have physical AI system for self-driving cars, for humanoid and there are also a lot of investments. So we have the great foundation and many young researchers who want to make a difference and we also have great need and investment. I think this is going to be a very exciting area and things can move very fast. I don't want to say that it will be solved in five years or 10 years, so I think it's still a long way. And more importantly, we also need to study how to best integrate these world models into the physical AI systems in a way that can really benefit them.

Kravitz: Right. And does that come through just working with partners out in the field, kind of combining research with application and iterating and learning?

Liu: Yeah, I believe so. I believe in suffering. So I believe that to hand in hand with our partners understand their problems is the best way to make progress.

Kravitz: For folks who would like to learn more about any aspects of what we're talking about, there are obviously resources on the NVIDIA site and of course the coverage of Jensen's keynote and the announcements. Are there specific places maybe a research blog, maybe your own blog, or social media channels where people can go to learn more about NVIDIA's work with world models and anything else you think the listeners might find interesting?

Liu: Yes. So we have a white paper written for the Cosmos world model platform and we welcome you to download and take a read and let me know how you know whether it's useful to you and let me know the feedback and we will try to do better for the next one.

Kravitz: Excellent. Ming-Yu, it's an absolute pleasure talking to you. I definitely learned more about world models and some of the particulars and the applications going forward, so I thank you for that. I'm sure the audience did as well. But you know, the work you're doing, as you said, it's early innings and it's all changing so fast. So we will all keep an eye on the research that you're doing in the applications and best of luck with it and I look forward to catching up again and seeing how quickly things evolve from here on out.

Liu: Thank you. Thanks for having me. It's been fun and I hope next time I can share more, you know, maybe more advanced version of the world model.

Kravitz: Absolutely. Well, thank you again for joining the podcast.

Liu: Thank you.