/mlst

60FPS Playable AI-worlds on Consumer GPUs [Overworld]

January 21, 2026

Machine Learning Street Talk

Podcast Transcript

Revision: b6b15414

Topic headings include retention indicators: ● above average, ● average, ● below average (vs. similar YouTube videos)

Contents

Overview

What if you could step inside your dreams and share them with others? In this fascinating conversation, we sit down with *Shahbuland Matiana* (Co-founder & Head of Research) and *Andrew Lapp* (Member of Technical Staff) from Overworld Labs to explore their groundbreaking new technology: *Waypoint 1* — an open-source world simulation model that runs on consumer hardware.

Unlike Google's Genie, which requires massive cloud infrastructure, Waypoint 1 is designed to run on your gaming PC. We're talking 3070s, 4090s, even Apple Silicon. This 2 billion parameter model generates interactive worlds at *60 frames per second* — and they're releasing the weights for free.

The Vision: Dreams You Can Record

Shahbuland shares a vivid lucid dream that shaped his entire research direction — a house floating in space, a circling dragon, a katana parry that cracked the floorboards beneath his feet. "This is the kind of thing where dreams can give you these really amazing fully immersive experiences, but there's no way to record them. There's no way to share them." That's what Overworld is trying to change.

How It Actually Works

The technical architecture is genuinely novel — a hybrid between a causal language model and an image diffusion model. Instead of predicting the next token like ChatGPT, it *denoises the next 256 tokens* representing each frame. Every sixteenth of a second, the model generates a new frame conditioned on all previous frames, your text prompt, and your controller inputs in real-time.

Why Privacy Matters Here

Tim raises a profound point: these simulations are extensions of our minds. When we imagine future scenarios, run mental simulations, or explore creative spaces — that's deeply private. The team agrees that running locally gives users ownership over their experiences in a way that cloud streaming never could.

We're Still Early

Perhaps most refreshing is their honesty about where the technology stands. Unlike LLMs, which have crossed from research into engineering, world models are still in active research territory. "Every other week, it feels like someone comes out with a paper that finds a way to make it 100 times faster."

This conversation covers the brain as a simulator, chaos theory in diffusion models, why fewer sampling steps reduce diversity, and the future of interactive entertainment. Whether you're a researcher, a game developer, or just someone

curious about where AI is heading — this one's worth your time.

Introduction & Overworld Demo

Andrew Lapp - Member of Technical Staff 00:00:00

I'm Andrew Lapp. I'm member of technical staff at Overworld Labs. I focus on pre training, post training, and inference.[1]

Shahbuland Matiana - Co-founder & Head of Research 00:00:07

I'm Shahbuland Matiana. I am a cofounder and head of research here at Overworld. I focus also on pre training, but mostly on the auto encoder side as well as helping with things all around for the pre training team.[2]

Tim Scarfe 00:00:18

Okay. Cool. So many in the audience, they would have seen that we did the launch video for Genie and that is this new technology that we're just trying to understand what the hell it is and how we could use it. It's really exciting and I guess you could call it a kind of continuous generative vision model that allows us to do a form of interactive entertainment, which is a little bit like dreaming or doing simulations of the world and actually interacting with it in real time.[3]

Andrew Lapp - Member of Technical Staff 00:00:45

Yeah. I think what they've done is really impressive. I think we're a little bit different in that we're a research project. We're letting the community run this on their own hardware. We're letting people explore the possibilities of Overworld and Waypoint 1 And we really wanna see what the community can do and figure out what all the use cases for this are. We have a few in mind ourselves. But, you know, making this open, making this public, making this accessible and runnable on consumer hardware, I think that's a big distinction that lets people, you know, figure out what is possible with this technology.

Tim Scarfe 00:01:16

Yeah. And that that is a big constraint. Right? So Google, they have it running on their TPU network. God knows how big the model is. It all just runs in the magical cloud and and you want to run it on consumer hardware. What do you mean by consumer hardware?

[1] Overworld Labs — Company The startup behind Waypoint 1, focused on consumer-accessible world simulation models.

[2] Google Genie — AI Model Google DeepMind's world model

[3] Waypoint 1 — AI Model Overworld's 2B parameter open source world model. Runs at 60 FPS on consumer GPUs.

Andrew Lapp - Member of Technical Staff 00:01:28

Consumer hardware that you can purchase your 30 nineties, 30 seventies, 40 nineties, 50 nineties, AMD, and we're gonna be targeting Apple silicon soon. So basically making it so you don't have to buy a $30,000 b 200 to run it. Making it so that anyone can run it on their own gaming hardware.

Tim Scarfe 00:01:45

Okay. Cool. And I think before we go into discussions, can you just show a demo of it just so that folks in the audience can see what it is we're talking about?

Andrew Lapp - Member of Technical Staff 00:01:52

So this is our overworld streaming demo. There's a few ways you can run Waypoint 1. You can run it locally on your own hardware. You can run it on overworld.stream, which is our service that you can go in your browser and try out. There's third party clients that allow you to run Waypoint 1. And we've really opened up the tooling and allowed people to stream it and run it however they want. But, yeah, this is our streaming demo. It allows you to enter a prompt. You can generate a world that you can explore, that you can play in. It's it's text experience basically. Real time interaction, the typical workflow for video diffusion models is you enter the prompt, you wait, and then you have a video. So this kind of disrupts that paradigm because you're able to real time enter control inputs, drive the video, drive the experience, and dynamically have anything you want to experience.

Core Technical Capabilities

Tim Scarfe 00:02:49

Okay. So this thing is running all the time. And initially, I think you said you can put a prompt in and you can just, like, generate a scene. Do I understand that you said that you also support this feature where you could kind of construct the scene interactively? So you build the scene and then you can say, I wanna have this thing and I wanna have this thing. You said the model is trained for that, but you're not supporting it at launch.

Andrew Lapp - Member of Technical Staff 00:03:11

Yeah. So the stack is available to run this. The model is trained with the capability to enter a prompt at the beginning and then adjust the scene either through controls or through a prompt that adds a new event, adds a new change to the

environment. So every frame that's generated is conditioned on the prompt and the controller inputs. So you can describe a change in the scene. That is not supported in this client. No. It's supported in our world engine inference library. Anyone can play with that. Anyone can test it. And, you know, you can make modifications to our open client as well. And we're planning on adding these new features such as in flight captions, we're calling them, that drive the scene. That isn't just to reiterate, that is not a capability at launch, but it's something that we're excited to be bringing soon.

Tim Scarfe 00:03:55

Okay. And you you said that you're supporting 60 frames a second or or you're aiming to do that. How much resolution is there? And, you know, how big is the model? And what are the kind of, like, performance constraints that you're heading up against?

Andrew Lapp - Member of Technical Staff 00:04:07

The model is this is our small model, actually. This is a 2,000,000,000 parameter model, and it runs at 60 FPS on a 50 90. The constraints of that really are you're just doing so much compute. You're generating 15,000 forward pass tokens per second. You're batching 256 tokens in a 4 pass, and you're and, basically, for some context, when you're you you convert an image into a series of 256 tokens, a 16 by 16 grid, and then a transformer model predicts the next frame based on all the historic history of the frames and all the conditioning, the text conditioning and the prompt conditioning. And that's a lot of processing. It's a it's a lot of work to do. So it it that's basically the main constraint, just how heavy the the the operations being performed on the GPU are.

Tim Scarfe 00:04:55

Yeah. Very cool. And at the moment, I think you said the context window is quite small, so it's about 2 seconds. But as we increase that, it will have a memory. So you could kind of, you know, look at something and you could look away from it and you could go back and it would still remember what that was. But is is how are you gonna kind of increase the length of the context window over time?

Andrew Lapp - Member of Technical Staff 00:05:13

Well, the big constraint is just a technological constraint that we've run into. And to alleviate that, we are applying We're basically distributing the sequence across GPUs when we're training so that we can actually have the full sequence. 30 seconds, 60 frames per second, that's 1,800 frames. You have to have a GPU process. And so we're we're sharding that sequence across GPUs so that the model can actually learn how to process such a long sequence.

Image Prompting & Experience Sharing Vision

Tim Scarfe 00:05:38

Okay. And at the moment, it's constructed with text prompts. But in the future, I I think you're going to introduce image prompts and and other forms of conditioning as well. Tell me about that.

Andrew Lapp - Member of Technical Staff 00:05:48

Yeah. So actually on day 1, we're supporting image to experience, image to real time streaming.

Tim Scarfe 00:05:54

Oh, cool.

Andrew Lapp - Member of Technical Staff 00:05:54

Prompting is just 1 way you can generate the experience, or you can just leave it unconditioned as well. Know, it's it's really the model is quite flexible. It's just a matter of implementing a client that can handle all these different functionalities. But on day 1, we are supporting image to video and text to video within the clients.

Tim Scarfe 00:06:11

And how much does it depend on the imagination and competence of the prompter? What I mean by that is, you know, we've all had this experience that Claude code is even it's like an amplifier of intelligence. So the more competent and the more you understand, the more you can specify things, the the more worlds you can create, the more things you can explore. Do you see a similar thing here? Do do you think that some people are really imaginative and they can come up with beautiful scenes and worlds? Or do you think that it's actually

quite democratized and and almost anyone can make this thing sing?[4]

Shahbuland Matiana - Co-founder & Head of Research 00:06:42

Unlike Cloud Code, there isn't a need for the 1 user to have to do everything themselves. We do think that as people explore the models, they're gonna probably want to customize the experience a bit to see what it can do and what kinds of things they can make for them. However, unlike Cloud Code and LMs, it is way more intuitive to share these kinds of experiences. You can imagine someone taking their prompts, taking their seed images, taking their instructions for how seed images are put in, things like that. Being able to basically share these directly. Imagine, you know, being able to go into a room, having a wall of all these experiences that people have made and just being able to step into 1 and experience that.

Tim Scarfe 00:07:18

How composable are these things? So I I guess, like, I'm imagining that level 1 of this tech stack is I've got the simulation machine. It's really interesting from a cognition point of view because I I think that our brain is like a simulator. And the reason why we have this amazing form of intelligence is because we can simulate things without, you know, direct physical experience and we can share those simulations with others through language. Language is pointers to the simulations. So that's level 0. That's very exciting. The next level is we have a social component. So we can like share simulations that we've created to other people and other people can compose them and extend them and so on. And then maybe the third layer of the stack is we have virtual reality and, you know, headset instantiation. So it'd be the killer application for virtual reality. But I I I guess the question is though, if you think about it, these simulations are constructed through a trajectory of conditioning. So it doesn't seem obvious how they could be composable and modifiable. How how do you see that working?

Shahbuland Matiana - Co-founder & Head of Research 00:08:19

I mean, the way that we're currently going about it, right, is that what as you as you play through an experience, right, and there's things that happen, there's events in the world, say, where you have these prompts existing that show certain things happening at certain points of the experience, We are currently primarily experimenting with text and being able to add text to the to the world. But the plan is in the future to also have this work for images, audio, and other kind

[4] Claude Code (Anthropic) — AI Tool Referenced as comparison for skill amplification - here the barrier is lower for creativity.

of media so that, you know, people that are creative, they might have doodles and art of, let's say, a boss or something. Right? And then be able to just inject that boss into the experience and have it spawned in. And then when they share this experience with others, you know, other people can fight this boss that they've created or something like that.

Lucid Dreaming Vision & Shared Experience

Tim Scarfe 00:08:59

When when we spoke earlier, you you were at pains to say that this is not necessarily a games engine. This is a form of interactive experience generation. And and you you were saying actually that you're quite interested in lucid dreaming and Yes. Some of those experiences in informed your your viewpoint here. Tell me about that.

Shahbuland Matiana - Co-founder & Head of Research 00:09:16

Yeah. So honestly, let me get specific. Let me talk about, like, an actual lucid dream I had, for example. And it'll make sense when I say that, you know, these are things that modern games can't really do. So 1 of the most cool experiences I've had my whole life, right, was a lucid dream. I was in this, like, house floating in space, and there was a giant, like, dragon circling the the house. And I I I I hear it detect me. It's, like, coming for me. I draw a katana from my, like, waist, and I parry the dragon's teeth as it goes to bite me. I feel a clang reverberate through my whole body. The floorboards crack beneath my feet, the window shatter around me, things like that. Right? And I I I I woke up and I was like, oh my god. That was awesome. Right? And this is the kind of thing where, you know, dreams can give you these really amazing fully immersive experiences, but there's no way to record them. There's no way to share them, And it's something that modern games just cannot hit because you cannot get to that level of immersion, right, where everything is is the world is bending around what you're doing. Right? I think that this technology, and fully developing it is the only way to get there. And this, like, my experience growing up with lucid dreams is kind of what's motivated me to pursue this this direction.

Tim Scarfe 00:10:24

Yeah. There's also this thing that, know, like our conscious experience and dreaming, it's it's subjective with a capital s, which means it's ungrounded. When I tell you about my dream and I say there was a blue swirly thing, you don't know what the blue swirly thing is because obviously there's no grounding physical experi-

ence between us. But here, this is a way of almost like curating a shared grounded experience, which means we can effectively share some of these amazing things together. But it also raises the question of I think that this might be different from Claude Code because in Claude Code, the cognitive wall is because some people can think in very high levels of abstraction and some cannot. But I have a theory that this is actually very grounded because we we we curate these experiences using possibly photos or videos or, you know, everyday descriptions of of our reality and that could actually Yeah. Democratize how we share these experiences.

Shahbuland Matiana - Co-founder & Head of Research 00:11:16

I I I I think it's there's an interface problem to it as well where it depends on how you present these people and how you allow them to share it. I think that 1 of the most important things going forward is going to be developing the social experience in a way that people actually are encouraged to, you know, share their experiences and try out their experiences, and let's say their friends have created or that others have created.

Open Source Origins & Platform Philosophy

Tim Scarfe 00:11:34

Do I understand correctly that you're sharing the weights for this? So is is it like a kind of semi open weights type situation?

Andrew Lapp - Member of Technical Staff 00:11:42

Yeah. The small model, which is what we've been demoing. It's a 2,000,000,000 parameter model, and that's gonna be open source. It's gonna be on a hugging face tomorrow of an AM at lunchtime.

Tim Scarfe 00:11:50

How did this whole thing come about?[5]

Shahbuland Matiana - Co-founder & Head of Research 00:11:51

Sora came out. I saw people using it for video generation, and I was like, well, this is really cool. It's like building these worlds, but you can't interact or step into them, but they look really good. So I figured that diffusion was was a much

[5] Sora (OpenAI) — AI Model OpenAI's video generation model that inspired Shahbuland to pivot from LLMs to diffusion research.

better direction to get to where I want to go, so I did a full pivot from LLMs into diffusion. And, you know, we're at stability back then. I spent few years mastering diffusion. Pretty great place to do that. And, you know, after the first role models came out, I was like it it ignited a passion within me, and I was like, okay. We should be trying to get diffusion as fast as possible and getting it into everyone's hands so they can experience it themselves.[6]

Tim Scarfe 00:12:24

It is and I think it it will be possibly the killer application, for AI, so I'm very excited about that. But 1 question I did have though is it it seems like you're you're going to, you know, you're trying to use local GPUs and you've got this kind of like local processing thing. But it seems to me like the direction of travel is platform, you know, platform occasion. Right? Eventually, think you said the kv cache on these, if I wanted to take a snapshot because it's not just the prompt, it's the conditioning trajectory, it's actually like the accumulated weights that I need to share. So it's several gigabytes. It's not something that I could easily share to people over the Internet. So the direction of travel is this will be running on the cloud somewhere. So why not start with the cloud?

Shahbuland Matiana - Co-founder & Head of Research 00:13:07

I'm sure I fully agree with that. I think that you can definitely transmit, maybe not the KV cache, but things like an image sequence or a seed sequence like that to someone's local computer to run it there. I don't really think that you are gonna be that heavily bottlenecked by by, you know, large data formats. Maybe for downloading the model itself, but beyond that, not really. Yeah. I also think it gives you a lot of privacy and, like, a lot more control if you can, you know, actually make things locally. It gives you free to, like, experiment with it too and and run different things.

Tim Scarfe 00:13:35

Yeah. I mean, the the privacy thing, this might be a bit of a vex topic to talk about. I hadn't really thought about that. But in in a sense, when we imagine, you know, like when I imagine future situations quite often we kind of have a a monologue in our heads and we say, well, in this situation I could have done this and I could have done that. And this is actually deeply private, so I can really see the the rationale for having this, you know, separated so it can't be shared with other people because this is like an extension of my mind. It should be very it should be very private. How how do you think about that?

[6] Stability AI — Company Where Shahbuland spent years mastering diffusion models before co-founding Overworld.

Shahbuland Matiana - Co-founder & Head of Research 00:14:03

No. I I I totally agree. I mean, it's 1 of those things where, you know, I I think if if if dreams could be recorded, right, and then everyone's dreams are recorded and shared with the entire world to see, it would be a bit dystopic. I think there is a level to which, you know, when you're when you're playing these kinds of experiences and you're exploring these worlds, there there should be a level to which that experience is your own. You know? There should be some sense of ownership over it. I think with streaming, you lose that.

Technical Architecture Deep Dive

Tim Scarfe 00:14:30

Just from a technical architecture point of view, so can you go into a lot more technical detail about how the how the model actually works and what sample you're using and just yeah. All all of the details.

Shahbuland Matiana - Co-founder & Head of Research 00:14:40

The thing this is built on is, like, is is having compression that can make the images much, much, much smaller, first of all. These models are not operating in, the actual pixel space. They're not you're not just generating raw frames. We have a lot of different directions we're going with this, but the 1 that way 0.1 is launching with at least is just kind of a pretty basic image compressor. It's an auto encoder that can compress 3 60 p videos into just, like, a small 32 by 32 image. Then from that, you know, you can compress videos into just 32 by 32 pixels. Right? Andrew can talk about what he actually does with those compressed videos.

Andrew Lapp - Member of Technical Staff 00:15:10

In a lot of sense, it's a combination of an LLM, a causal LLM, and image diffusion model. So in an image diffusion model, you're given a set of patches, and you figure out how to denoise them. And you condition these denoise patches on text usually. But in our case, you are generating a new patch, a new a new frame, I should say, a new series of patches every sixteenth of a second. And each of these each of these frames are you know, it's just like an image diffusion model, but they're conditioned not only on the text, but also on controller input from the last sixtieth of a second from the last frame in all proceeding frames. So what that what you end up with is just kind of a standard transformer LLMS, except instead of generating the next token, you're denoising

the next 256 tokens. You know, it is a standard feed forward transformer. It has attention in MLP like a transformer. It conditions using cross attention between between the current activations and the text embeddings, the controller inputs are our own home baked controller input embedding we're calling it. So that's kind of like the gist of our architecture.

Tim Scarfe 00:16:17

Okay. So do I understand correctly that it's a sequential architecture? So you're processing or or you're kind of like decoding 1 frame on its own and then you move to the next frame or do you have some kind of cross frame processing going on?

Andrew Lapp - Member of Technical Staff 00:16:30

We only generate 1 frame at a time. If we generated more than that, then you would kind of have a lag in input in in controller inputs. Right? So if you're generating 5 frames at a time, 8 frames at a time, you then have to wait for those 8 frames to be generated before it's actually responsive to the controller input. So for that reason, we're kind of constrained to generating 1 frame at a time.

Shahbuland Matiana - Co-founder & Head of Research 00:16:48

Yeah. Playability is super important for us. I think that when you start to introduce latency for the player, which come from streaming, if you're not doing streaming efficiently, as well as from, you know, using a traditional autoencoder. Like, most of these video diffusion models, they they use a temporal autoencoder, which basically means that they compress every 4 frames so that they become 1 latent frame. So their diffusion model underneath the hood doesn't actually need to generate that fast. In fact, if you are, let's say, trying to aim for 24 FPS, right, and you have a 4 x temporal compression VAE like, you know, WAN or HODYAN, those kinds of models, then your late your laden model only needs to generate at, let's say, 6 FPS, and that 6 FPS can be upsampled to 24 FPS. However, you would not be able to take user input every frame. You'd be taking user input once every fourth frame, which can add pretty bad latency. You'll see with a lot of other world models, they'll have, you know, basic WSD controls and arrow keys to look around and things like that because you can't do proper high frequency controls like mouses or eventually, like, you know, VR headsets looking around. If you are only taking controls once every fourth frame at 6 FPS and if there's a delay of worth, let's say, 500 milliseconds, it's a no it's a no go.[7]

[7] HunyuanVideo (Tencent) — AI Model Video diffusion model mentioned as example of 4x temporal compression VAE approach.

Optimization, Distillation & Future Outlook

Tim Scarfe 00:17:56

And what kind of sampler are you using? And are you doing the same amount of computation every time, or do you have some kind of adaptive decoding?[8]

Andrew Lapp - Member of Technical Staff 00:18:04

So we we don't have any adaptive decoding. We're using the same 4 step flow matching Euler sampler. So basically, we're we're doing a rectified flow model. So it's gonna just generate 4 step diffusion, 4 step denoising, same trajectory every single sample.

Tim Scarfe 00:18:22

And can you just explain just just for the audience who don't understand how this process works? Can you know, like like a, you know, the rectified flow model and so on. Can you just explain how that works?

Andrew Lapp - Member of Technical Staff 00:18:29

Basically, in our rectified flow model, you are sampling a random point of noise in space and there's a clean ground truth somewhere else in space. And the rectified flow model predicts the vector that'll get you closer to that clean point, that ideal point, the point that you're trying to generate conditioned on the inputs. But it's not always right there. So you have to have multiple steps so it can kind of correct and kind of gravitate towards the ideal denoise frame. So, you know, it's it's multiple transformer passes. Each of them is predicting what the vector is. They'll move you from the pure noise input to the partially noise input to the clean output.

Tim Scarfe 00:19:08

And can you tell me about some of the trade offs here? So, know, there are all of these different parameters you set on the architecture, like, you know, how much decoding do I do? What do I set the parameters to and so on? How did you go through that engineering process?

[8] Rectified Flow Models — Paper/Technique The diffusion technique used in Waypoint 1 - predicts vectors from noise to clean output.

Shahbuland Matiana - Co-founder & Head of Research 00:19:18

The 1 thing that has been found kind of in the literature for division installation is that very often the thing you're actually sacrificing when you reduce step count is more often than not diversity as opposed to actual quality. The here's the interesting thing. For image diffusion, this is a pretty big deal. If people want very diverse and interesting images for every prompt they give, they'll often use a higher step count, so they won't use distilled models at all because it lets you retain that diversity. However, for autoregressive diffusion models, it's kinda like this cross hybrid that we're using between an autoregressive transformer and a diffusion model, it doesn't really matter that much because your your conditioning is more than just the prompt. Your conditioning is the previous frames. Your conditioning is the control inputs. Your conditioning is the text. All these things put together. So the loss of diversity that these distillation methods give you doesn't really matter anymore. What we do notice is that it seems to be the case that you can drop the step count to 4 during distillation without really losing any quality. It's only when you start to go to 3, 2, or 1 where you see big sudden jumps in it. I'm of the opinion that, even though people say in the literature that 1 works, it's a bit of a strange thing because if you're doing 1 step diffusion, it's not really diffusion anymore. Right? It's at that point, basically, GAM because you're going directly from noise to an image. It's it's a bit weird. But 4 seems to be the sweet spot where you're not really losing any quality. There is, of course, the trade off for speed. Right? 4 diffusion steps, 4 forward passes, 5 actually, because of how the way that our setup works. That's a lot of added latency. Right? So even if your model is running at, let's say, a 100 FPS, if you do 5 steps, it's gonna be 20. Right? So you guys gonna drop down to 20 FPS. So there's a bit of a balancing act here. We think that at at 2 diffusion steps, you can still get a pretty good level of quality in a lot of cases, especially for these bigger models without needing to sacrifice too much speed.

Tim Scarfe 00:21:08

Yeah. So so you you said that in traditional models, more step count means more diversity. Yes. And what's the what's the intro like, why why is that the case?

Shahbuland Matiana - Co-founder & Head of Research 00:21:16

When you're when you're doing these long trajectories, right, the paths are very, very chaotic. Even though rectified flow is technically trained to do a straight line from input to output, that's not really how it works in practice. In most cases, the path that you take is gonna be very curvy and wiggly, and the place that you end up is very chaotic depending on where you started. I've actually done some experiments with optimized versions of these models where I can play around with the input starting noise in a 2 d space, and I find that as, you know, if you move your cursor around in the space and you would give it different starting noise, the image you end up with is very different. It's kind of, honestly, an application of of of chaos in the area of sense. Like, the starting conditions really have heavy effects on where you end up in the end. Right? But when you do a lot of these distillation methods, they are analogous to kind of forcing the line to be straight. So like I said, it's normally wiggly. Right? But when you try to do it in in, like, 1 or 2 steps, you're basically forcing it to be straight. A straight line often ends up at the same place or at least very close. So you do sometimes lose that diversity when you do low step sampling, especially for just raw text image with nothing else in the mix.

Andrew Lapp - Member of Technical Staff 00:22:22

The safest place for a rectified flow model to point to is the middle of the data distribution. And once you're in the middle of the data distribution, you if you're in a 2 step model, you point to whatever is safe to point to, basically, if that makes sense. So you're basically often getting very close to mean the the mean of the samples, the mean of the model distribution.

Shahbuland Matiana - Co-founder & Head of Research 00:22:41

And and funny enough, if you if you don't do any proper diffusion distillation methods and you try to do a 1 step or 2 step generation with a rectified flow model, so without any distillation, you'll actually see that it just it just mode collapses. It it mode collapses intentionally. You'll see it generate something blurry that looks like the meat of your dataset.

Tim Scarfe 00:22:57

Yeah. I wanted to talk about that. So so many folks at home, they would have played with stable diffusion models and they probably had the experience of, you know, you have all of these things that you can play with. Right? So, know, am I gonna use LCM? Am I gonna use Caras? Am I gonna use Euler? You know, like, there's this config parameter, you know, which is about, like, how much should it pay attention to the prompt and so on. And I I I guess intuitively people find that there's a Goldilocks zone for these parameters where you can't really move too much because otherwise you get mode collapse and the image diverges. Yeah. So do do you envision a world where you're you're just setting these parameters and and they're the same for everyone? Or or would they reasonably depend on the type of hardware people are using and the type of images they're generating? Would you ever foresee the users themselves changing these parameters or do you think they should just be fixed?

Shahbuland Matiana - Co-founder & Head of Research 00:23:40

When you're doing distillation, you do lose control over a lot of these things. Just as a basic example of that. Right? 2 axes that I think are probably the biggest things that people use for controlling these things is CFG scale and the actual scheduler that you use during during inference. Right? But here's the problem. When you do DMD or you most of these methods of of, diffusion distillation, you lose control over the schedule. For for DMD, for example, in our setup, at least, we we pick specific noise levels that the student isn't trained on. So for example, the student might be trained on only let's say this is what I've asked. It's actually a bit of an example. Let's say the student was only trained on, like, 1, 0.75, 0.5, 0.25 for 4 step. Then if you were to feed it, like, a sample, let's say, with 0.8 noise, it just wouldn't be used to that. Right? You can't just give any schedule you want during inference because during distillation, you fix the schedule. You also fix the guidance scale because it's distilled, you know, to do classifier free guidance built into it.[9]

Andrew Lapp - Member of Technical Staff 00:24:38

Yeah. Otherwise, you need to do a forward pass for each conditioning vector, each conditioning signal, and an unconditioned forward pass, which is really expensive. So not only would be be performing all our denoising steps, so we'd be having to do them 4 times. So do you you know, basically, you bake that in and say there's only 1 value for the conditioning signal. You don't need to do unguided and guided. You hard code the CFG during distillation. And then you end up with

[9] DMD (Distribution Matching Distillation) — Paper/Technique Distillation method discussed for reducing diffusion steps while maintaining quality.

a model that is, you know, might need to be post trained for a different setting if you want a different setting, but it's 4 times as fast.

Tim Scarfe 00:25:09

Do you think that it is bottlenecked on hardware or algorithms? I mean, do you guys know this far better than I do because you're studying the state of the art of all of these different algorithms. Do you think there could be some kind of breakthrough that could just unlock this and just really make it tractable on consumer hardware?

Shahbuland Matiana - Co-founder & Head of Research 00:25:24

It's there there's like there's like a point with LLMs where it broke past this threshold where it stopped being mostly research and started being mostly engineering. Don't think we're at that point for role models. Don't even think diffusion itself is at that point. You know, I think that there is still a lot of ground to cover. There's new papers coming out all the time about ways to speed up diffusion training, ways to make smaller models more efficient, ways to do, say, distillation. There are so many axes of improving these models and making them smaller that it would be nonsense to make assumptions like, oh, this will never run locally. Oh, this will never run on a phone. Oh, this will never run on then then it doesn't it doesn't make sense because every every other week, feels like someone comes out the paper that finds a way to make it a 100 times faster. You know?

Andrew Lapp - Member of Technical Staff 00:26:03

Yeah. In the short term, we have the capability to reduce the step count without reducing without reducing fidelity. And we also have the opportunity to quantize the model with only a very marginal decrease in fidelity. So those are our 2 targets to immediately improve the speed 3 x. And there's also a variety of strategies that allow us to scale up the model without losing a substantial throughput, and there's also a variety of strategies that are basically targeting improving model quality for the size the model is. So there's it is a very dynamic research space.[10][11]

Tim Scarfe 00:26:35

And your current

[10] Overworld GitHub — Code Repository Official GitHub organization with open source world model code, inference engine, and tools.

[11] Overworld Discord — Community Official Discord community for Overworld users and developers to collaborate and share creations.

References

[1] **Overworld Labs**
https://over.world/
Company The startup behind Waypoint 1, focused on consumer-accessible world simulation models.

[2] **Google Genie**
https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/
AI Model Google DeepMind's world model

[3] **Waypoint 1**
https://huggingface.co/Overworld
AI Model Overworld's 2B parameter open source world model. Runs at 60 FPS on consumer GPUs.

[4] **Claude Code (Anthropic)**
https://www.anthropic.com/claude-code
AI Tool Referenced as comparison for skill amplification - here the barrier is lower for creativity.

[5] **Sora (OpenAI)**
https://openai.com/sora
AI Model OpenAI's video generation model that inspired Shahbuland to pivot from LLMs to diffusion research.

[6] **Stability AI**
https://stability.ai/
Company Where Shahbuland spent years mastering diffusion models before co-founding Overworld.

[7] **HunyuanVideo (Tencent)**
https://github.com/Tencent/HunyuanVideo
AI Model Video diffusion model mentioned as example of 4x temporal compression VAE approach.

[8] **Rectified Flow Models**
https://arxiv.org/abs/2209.03003
Paper/Technique The diffusion technique used in Waypoint 1 - predicts vectors from noise to clean output.

[9] **DMD (Distribution Matching Distillation)**
https://arxiv.org/abs/2311.18828
Paper/Technique Distillation method discussed for reducing diffusion steps while maintaining quality.

[10] **Overworld GitHub**

https://github.com/Overworldai
Code Repository Official GitHub organization with open source world model code, inference engine, and tools.

[11] **Overworld Discord**
https://discord.gg/overworld
Community Official Discord community for Overworld users and developers to collaborate and share creations.