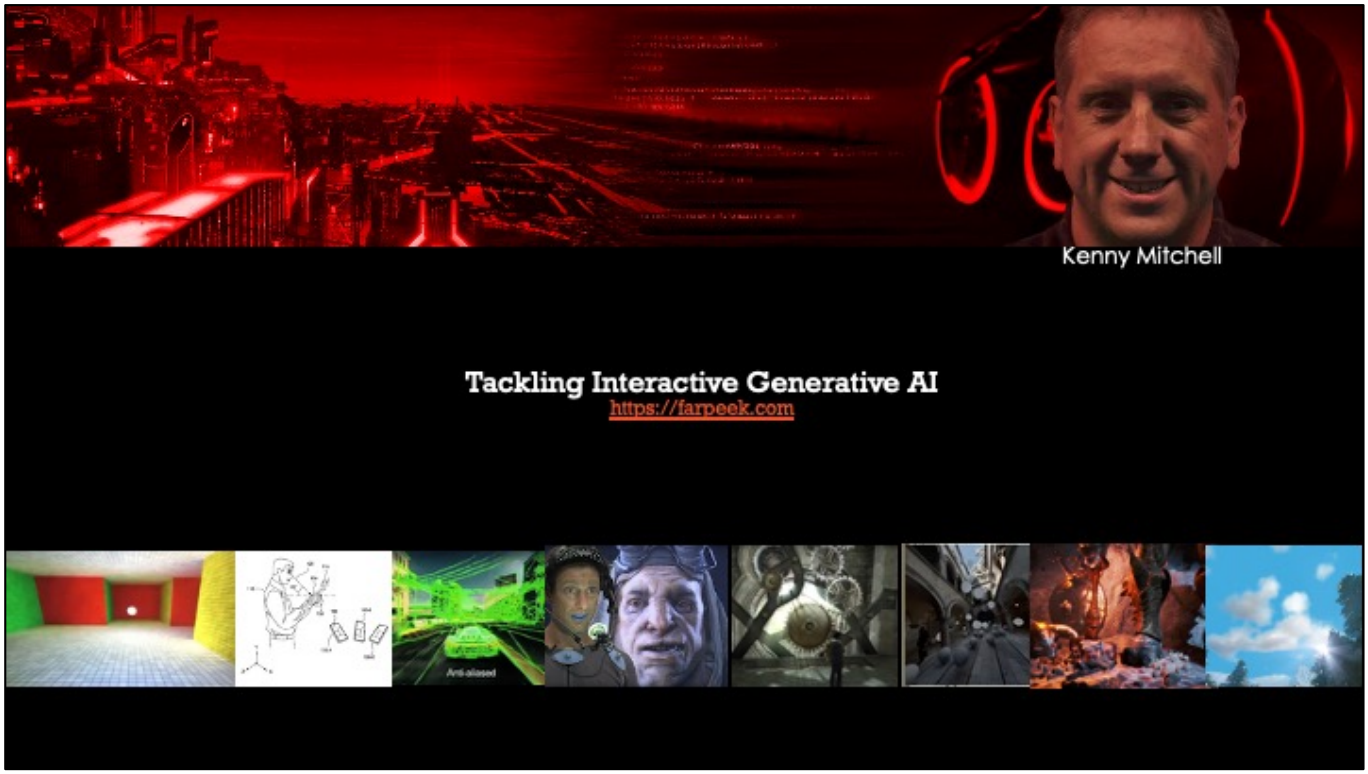


<click to play vids>
Hi, thanks very much for that kind introduction and I'm really appreciative
to have the opportunity to speak with you all today
on some research developments that completed last year.

For a bit of context , most of my career efforts have been around empowering creativity
and making media production faster and easier to get ideas on the screen,
interactively for games, for movies and even physically into themepark attractions,
Contributing to around 50 patents for Disney so far,
it's always been about opening myself to take risks personally on pushing technology but strategically within a responsible setting.

This includes leading rendering visual features for millions of video game creators and players
for Roblox
and for Harry Potter fans with Electronic Arts.
Mostly though, I've performed this 'enabling others' role in the last decade for Disney Research
where I established a lab of 20 artists and researchers in Edinburgh in Scotland as part of Disney Research Studios.
Today though, i'm presenting more recent research works
as I finish up being a part-time professor for Edinburgh Napier University where

led a consortium of universities on the project called, CAROUSEL
which delivered immersive low-latency technology for dancing online using AI for low-latency motion prediction,
and full body conversational generative AI avatar agents.

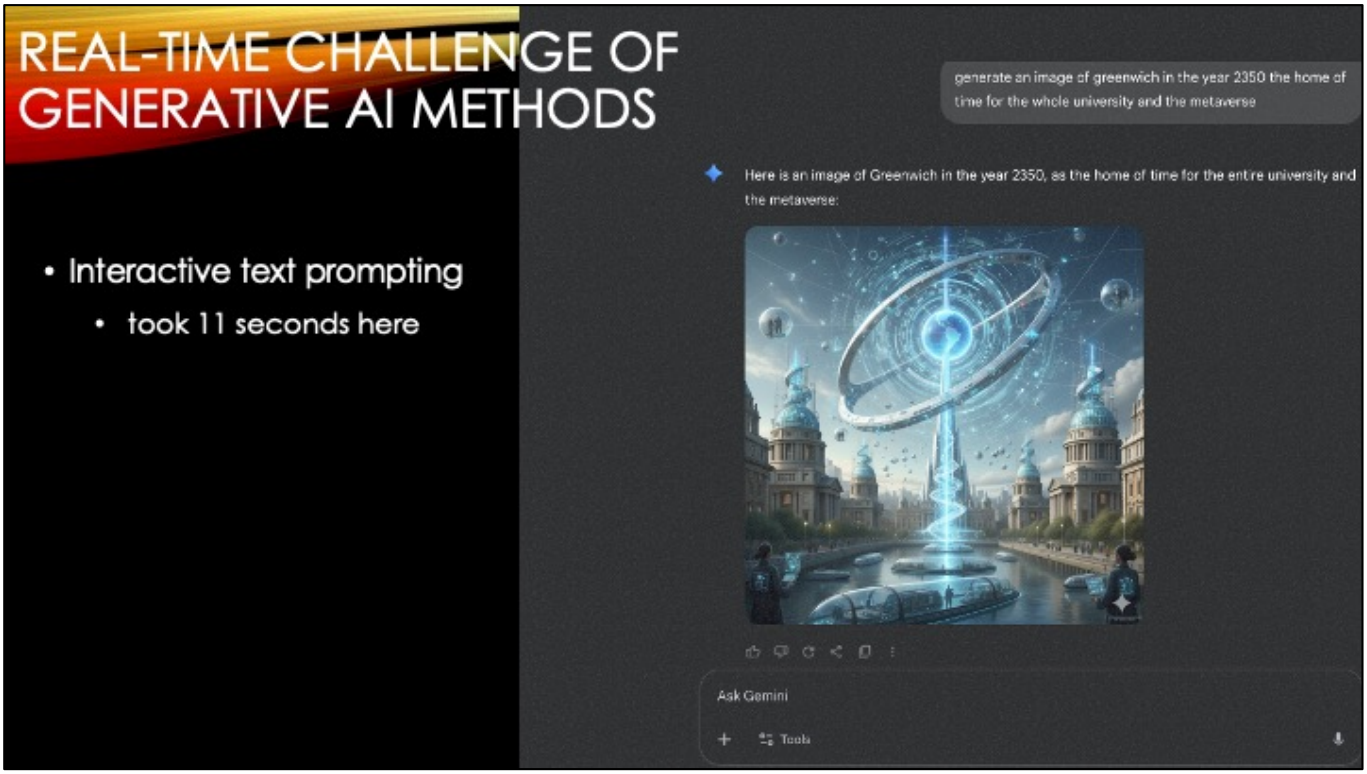
This selection of visuals here shows off a bit of those other works from
introducing real-time global illumination on iPad,
gaze dependent 3D holograms on iPhone,
introducing PBR for the Disney racing game Split Second: Velocity,
real-time digital acting for ILM and Star Wars,
real-time god rays for Harry Potter,
light field rendering for Animation Studios,
foveated rendering in Unreal Engine for Imagineering,
scalable volumetric clouds for Roblox, and more.
But let’s go ahead and move on to todays’ topic of interactive generative AI.

So, I'll begin here with an overview and some motivations.
firstly, it's a good exercise to recap why it's important to have fast responsive interactive tools and systems.
Many creative production workflows take a process of iteration, from rough initial ideas,
trying out sketches on paper and then refining as the idea becomes more concrete, directed, and honed
according to the artist's vision until it's ready to ship to the audience.
But, there's a crucial constraint here, if a single edit or correction takes seconds, minutes or even hours,
then the whole creative iteration and flow of thoughts gets stalled
and commonly most of our artistic experience today is staggering through this stop start cycle of interruptions
while we wait to see the outcome of our creative intents at every step.
To bring this key idea of low-latency even more profoundly, take the iphone for example,
when iphone transformed our lives almost 20 years ago now,
it not only focused relentlessly on a compact design unencumbered by a stylus or clicky buttons taking up space on the screen, but on low-latency.
When you swipe left or right it followed your finger's moves live,
apps responded faster than ever before to touches on screen,
the whole design philosophy revolved around making the user experience feel incredibility fluid and instantaneous.

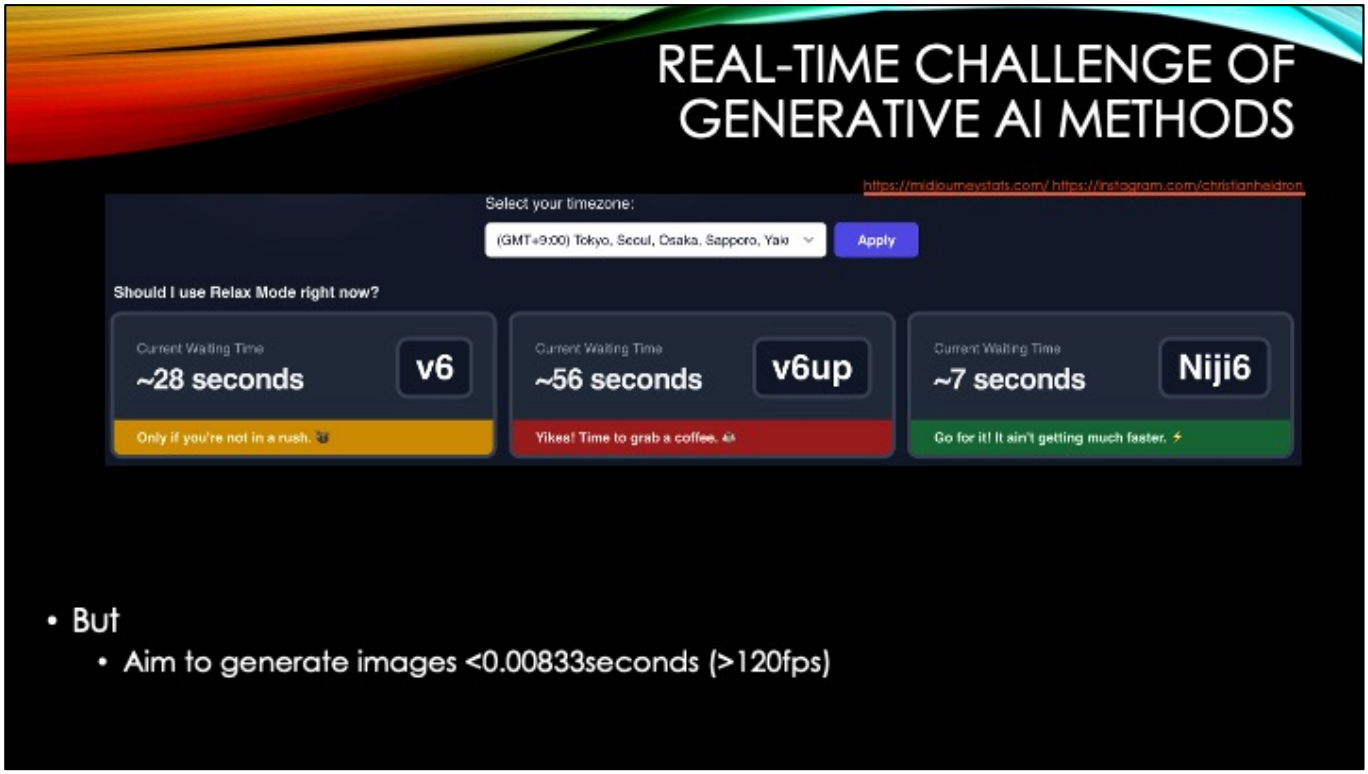
And this I'd argue the key reason mobile phones have become so pervasive and almost extensions of ourselves in the past two decades.
Without being interactive and fluid any new technology tends to flounder as iphone's competitors found,
and indeed as I believe we'll find this same progression in our new transformational wave of generative AI.

To set the ultimate challenge, I really proposed an impossible vision here,
Where we want to be able to generate any content in real-time live
and be able to control that content as meticulously as we would making a movie or building a video game
and just to make things a little more ambitious let's say our ultimate vision is of a real-time generative AI metaverse,
where were sharing a creative world online in virtual reality,
And crucially we need that interactivity to be real-time in VR at 120frames per second,
not just at interactive rates in a digital content creation application like blender or photoshop

And one note to mention with research it's often encouraged to think about the bluesky challenges
which aim to become relevant between now and 10 years in the future,
so just to explain some ideas might be pretty radical and forward looking
and this approach aims to try out a few ideas to better understand where limits and opportunities might actually land in practice.

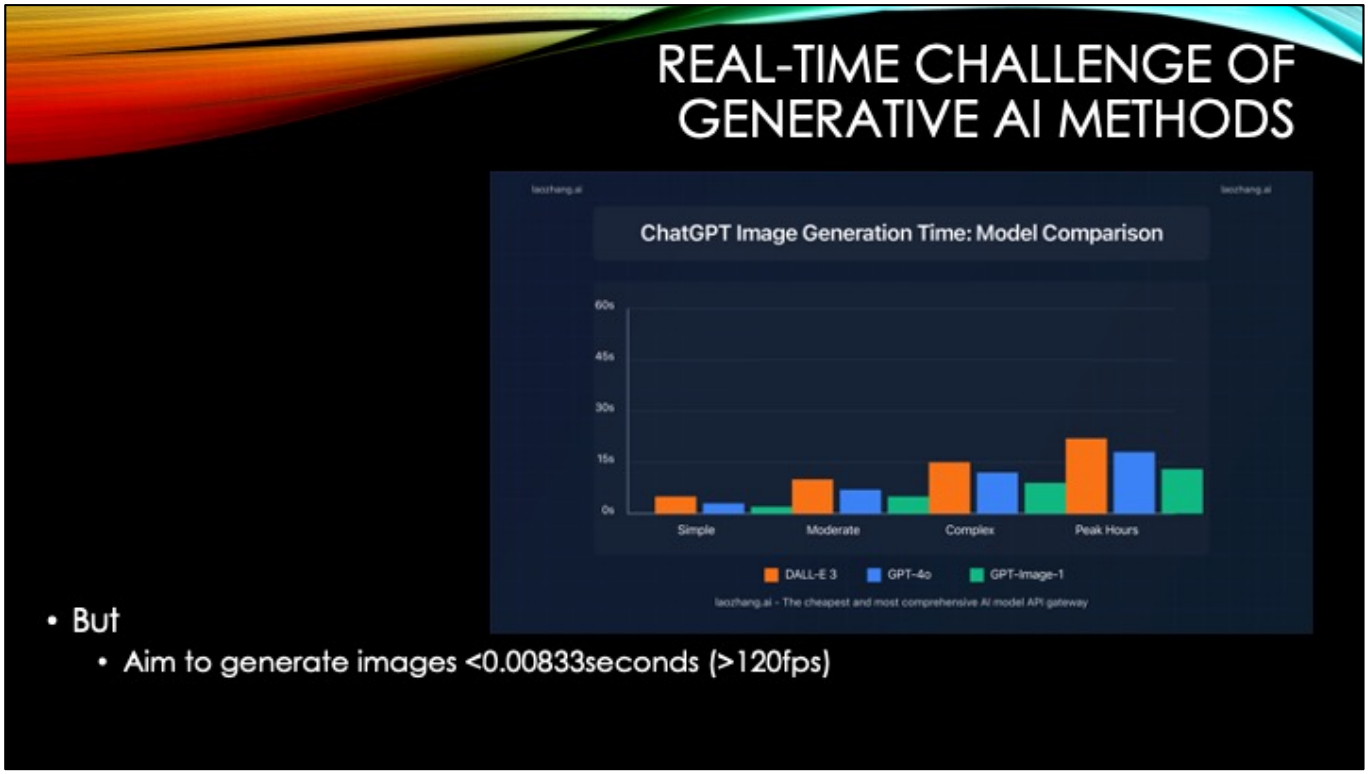


Here's google deep mind's Imagen Nano Banana model in practice last october, with a slightly mis-worded prompt I generated for a similar talk I gave in London near by the global time center in Greenwich
and it produced this very nice image in around 11 seconds in the cloud,



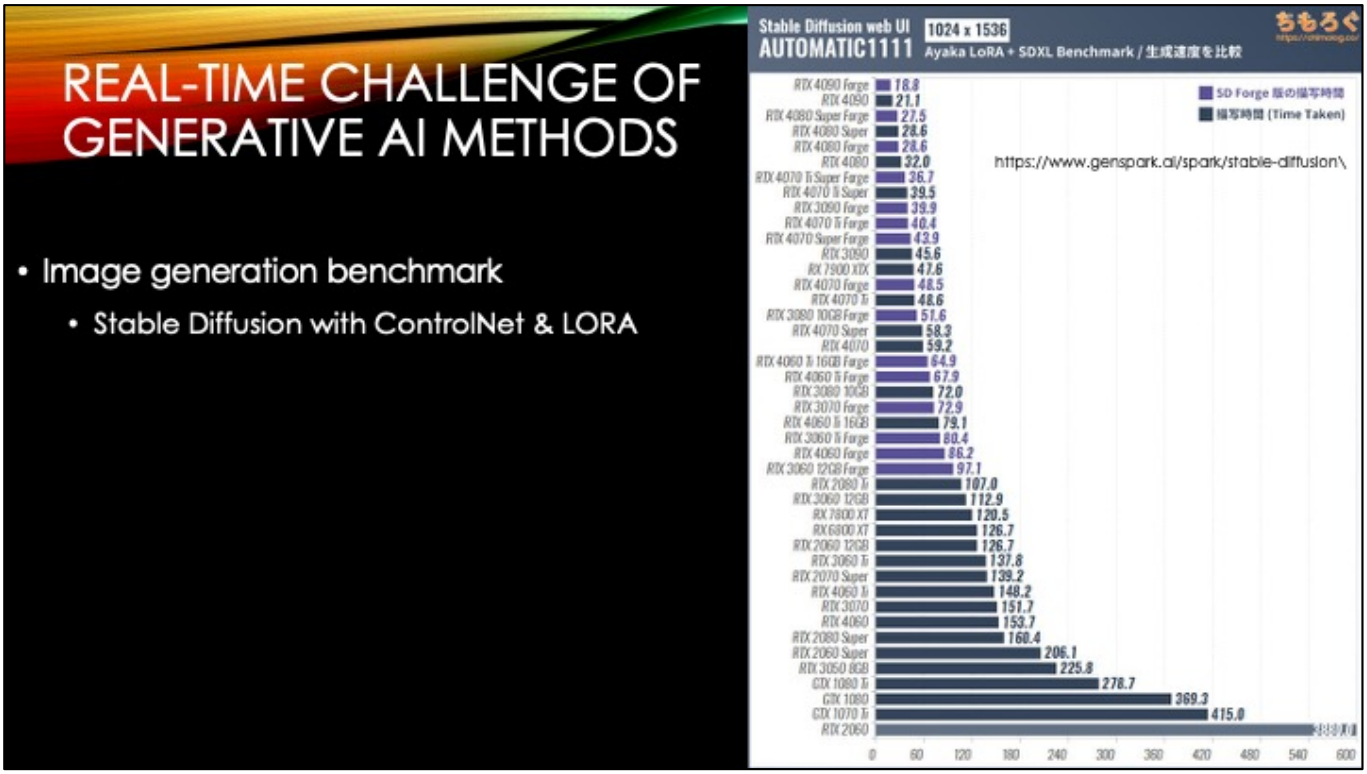
Diving in a little more on how long things take,
in order to be interactive and even contemplate achieving metaverse-like immersive interaction in virtual reality it's fundamental to be able to update generated image frames with low latency at over 120 frames a second for each eye to lower the chance of motion sickness and provide a smooth user experience.

The top graphic here shows older timings of latency for MidJourney image generations,
where they have powerful cloud compute performance accessed via their discord bots.
also we see the quality to latency trade off they're users have been making to retrieve images with alternative delays for better or worse quality in the results.

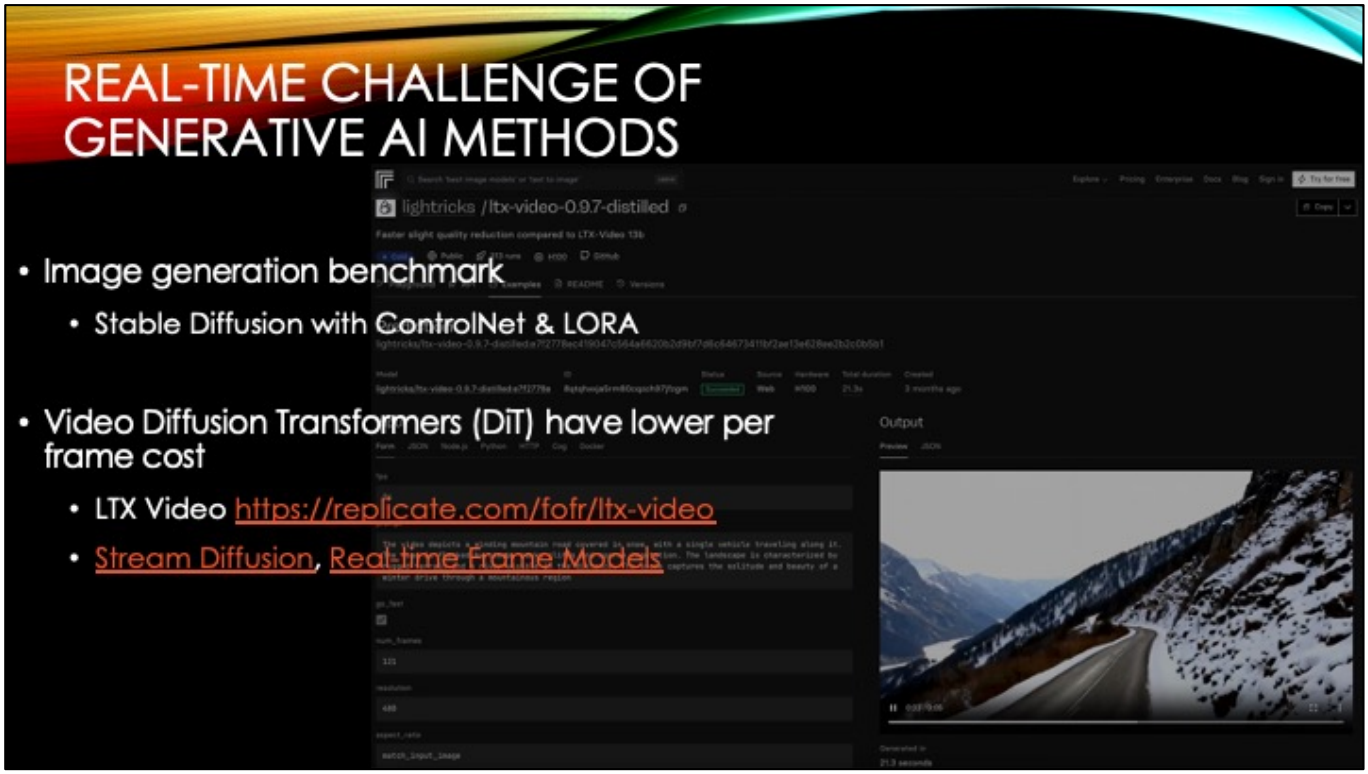


And a more recent plot for ChatGPT here, with alternative quality levels and models, shows a similar trade off folks can make.
For example, trading settings according to whether it's early concept development images or previs imagery getting towards production ready, but generally still taking around 10 seconds.
and in practice right now with increasing demand the figures often don't meet these plotted numbers recorded a few weeks ago.

And of course, this isn't a live interaction use case like games or VR that image generation services target,
but hopefully this gives more sense of typical round-trip times to be expected of static image generation solutions.



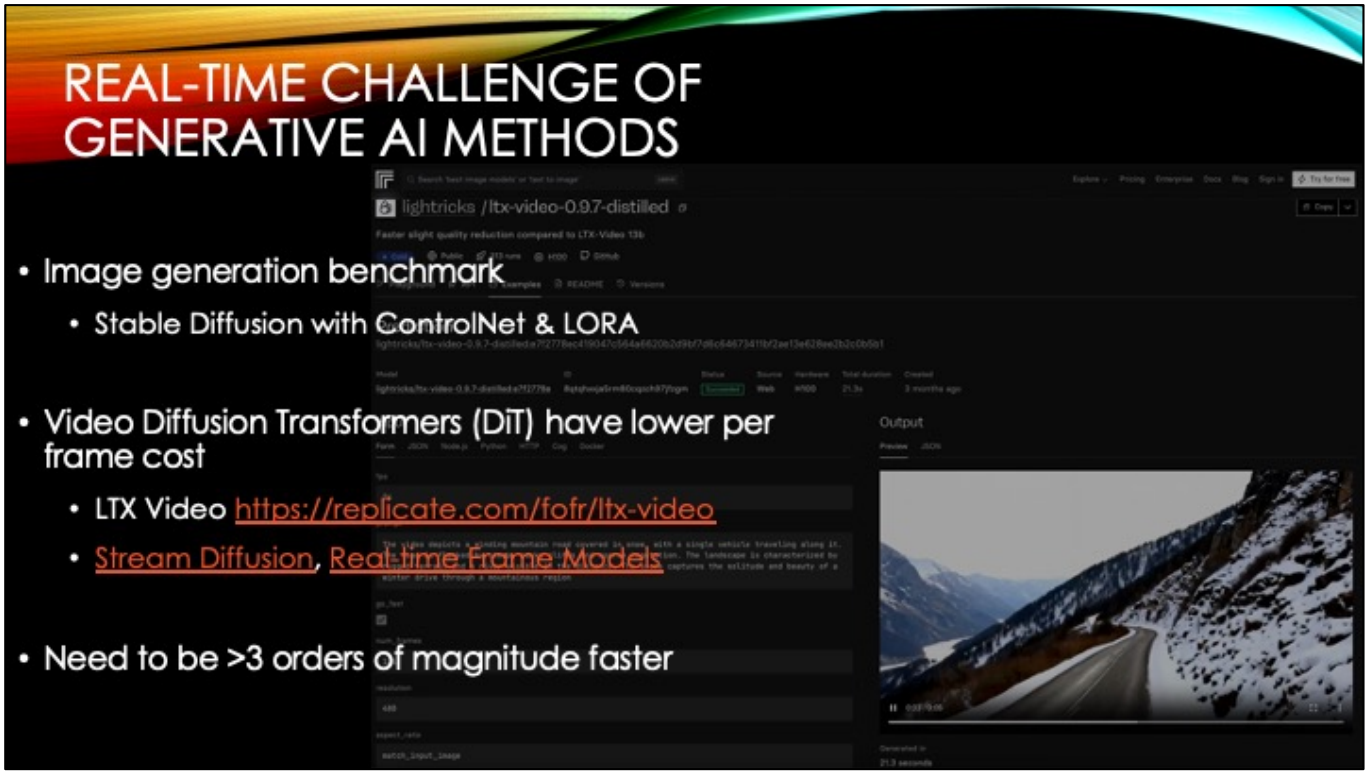
This is another nice example to get an idea of performance timings for generative AI,
but when computed locally on your own PC without cloud network latency,
so we have around 20 seconds on an admittedly hard benchmark case
using an advanced ControlNet configuration
with a LORA stylization adaption of a Stable Diffusion XL (SDXL) model
And this is running on a high-end consumer RTX GPU,
Which is pre-Blackwell benchmarks and not including H100s



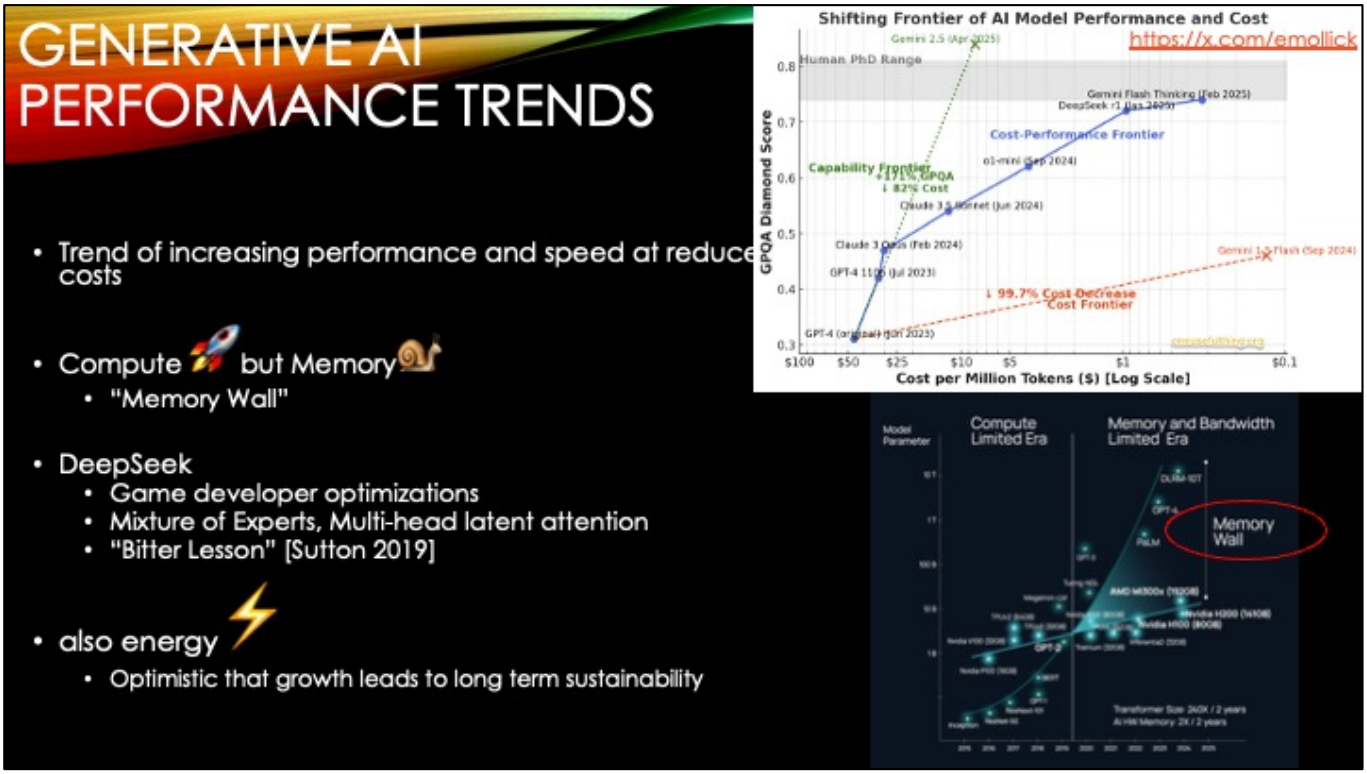
Certainly though, a few research approaches have better per frame cost such as video diffusion transformers like LTX Video
which technically can generate frames faster than movie frames can be seen,
That's 24fps for text, image or video to video cases.

Which is great, but either the whole video sequence is computed end-to-end
or you have the whole source video supplied up front,
so there's still a gap to provide real-time interaction with genAI content freely.

Even given the very latest Real-time Frame Model
which generates next frames for 3D environments in real-time
with a similar optimized approach
needs a 20k$ H100 GPU right now



And so regularly right now we need at least 3 orders of magnitude improvement in processing times for live interaction
with full generative AI content adapting dynamically to your design inputs



It's not at all bad news though, as folks maps the trends of improving performance progression
and importantly the persistent trend of getting more out of every compute dollar, as this plot in the top right shows

While compute performance leaps ahead, actual memory performance is not advancing as quickly,
And this is characterized as the Memory Wall limits in both hardware bandwidth and capacity.

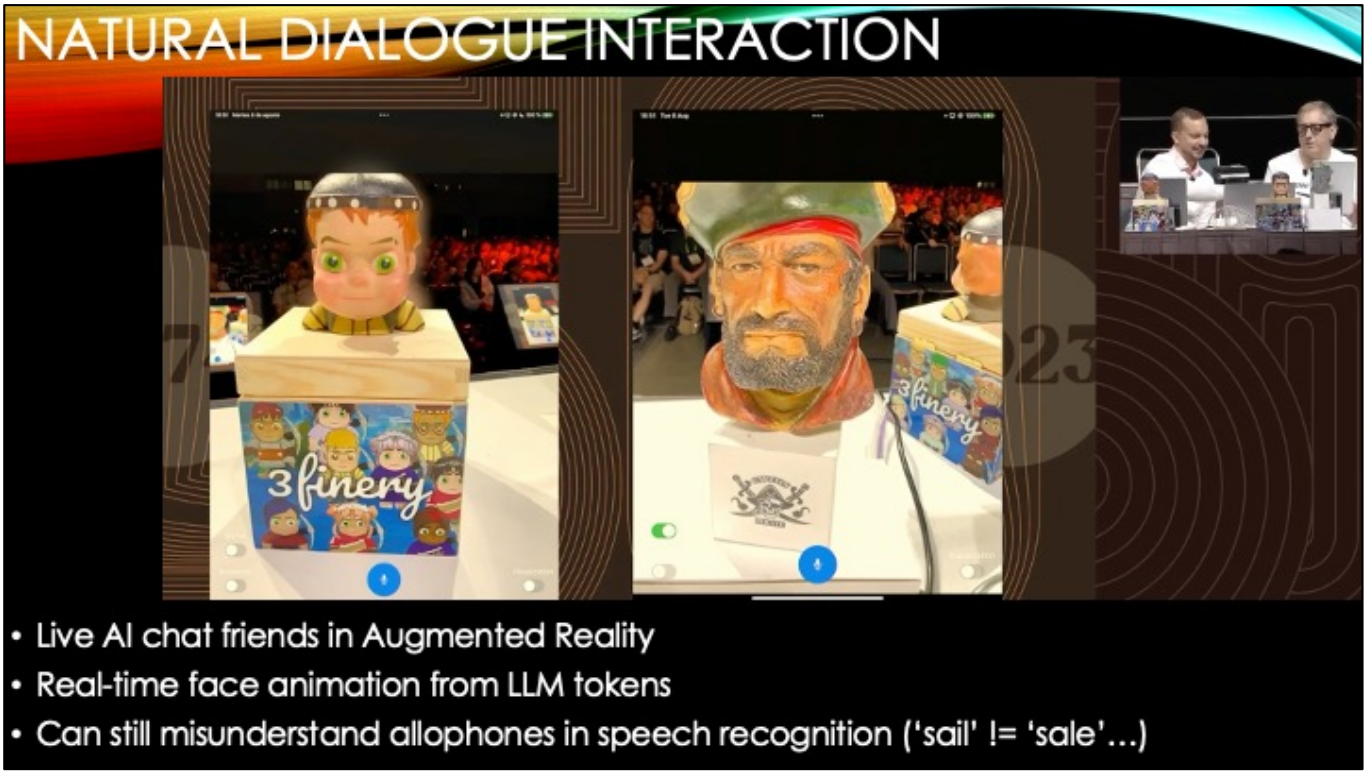
This Memory Wall term was established before developments like DeepSeek in late 2024 that delivered a leap in performance, but is still a key cost
For me, I consider DeepSeek in a way applied some optimization schemes actually already quite familiar to game developers,
using quantization and bare metal coded GPU CUDA approaches,
And that couped with memory local architectural changes like mixture of experts, multi-head latent attention
to really push on breaking down this memory wall
and not address all the data space all the time that prior LLMs tended to suffer from.

Also important to say is the something called the 'Bitter Lesson' from computer vision

where perennially reducing compute costs rendered many traditional computer vision methods
that took decades of development (like features descriptors) are now defunct.

And we can't talk about AI trends without considering the energy costs involved, which I see it maybe differently.
Right now, it's true there's an explosion of energy use in compute farms and that's super concerning,
but I'm also reminded of the 'bitter lesson' and similar progressions,
where the mobilization of energy generation is also likely to lead to greater energy production efficiency
and more investment into sustainable energy generation,
so I actually see this trending positively also,
not without cost in these years,
but I'm optimistic of this in the long term and new optimizations of large model research
like hierarchical and adaptive models and more information efficient representations of tokens.

So, while we still have typically 10 seconds per generative AI image
and a couple of minutes or so for video generation
there's a number of baby steps we can already take
towards our impossible interactive generative AI metaverse research challenge



And in the first place we might start by making this more human centric and natural through speech
and turns out we're already quite well developed on integrating LLMs and speech recognition with live visual interactions.

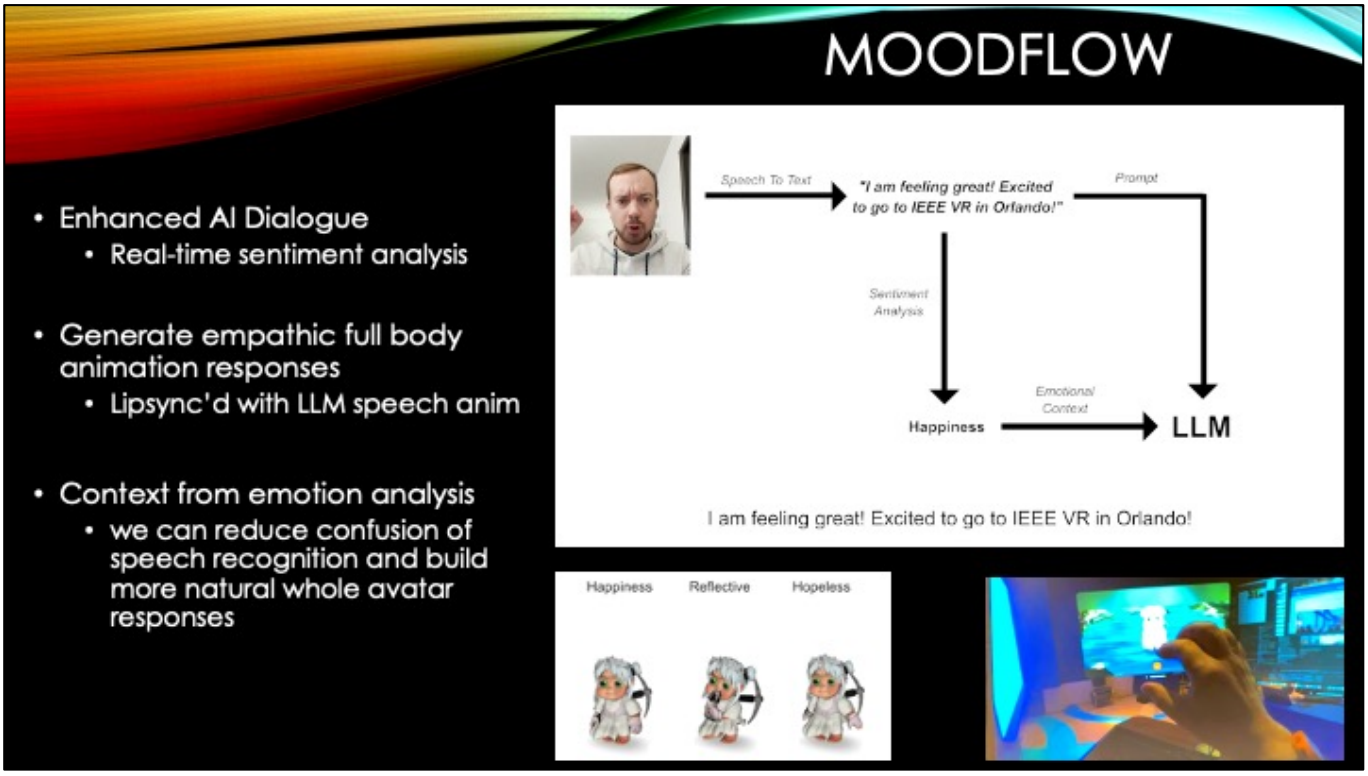
For example, we already mapped the tokens of text responses to facial animation blend shapes
and animated with warping of a live video stream registered to these 3d printed characters in augmented reality

let's take a look at this video of our 2023 SIGGRAPH Real-time live feature where we presented conversational generative AI in augmented reality
<play>
For sure this was true to the tradition of live demo glitches in front of 1000s of people in the audience,
the visual appearance of the 3d printed animated figures worked out pretty seamlessly,
But the lip sync animation could certainly use more advanced rigging and co-articulation of viseme sequences,

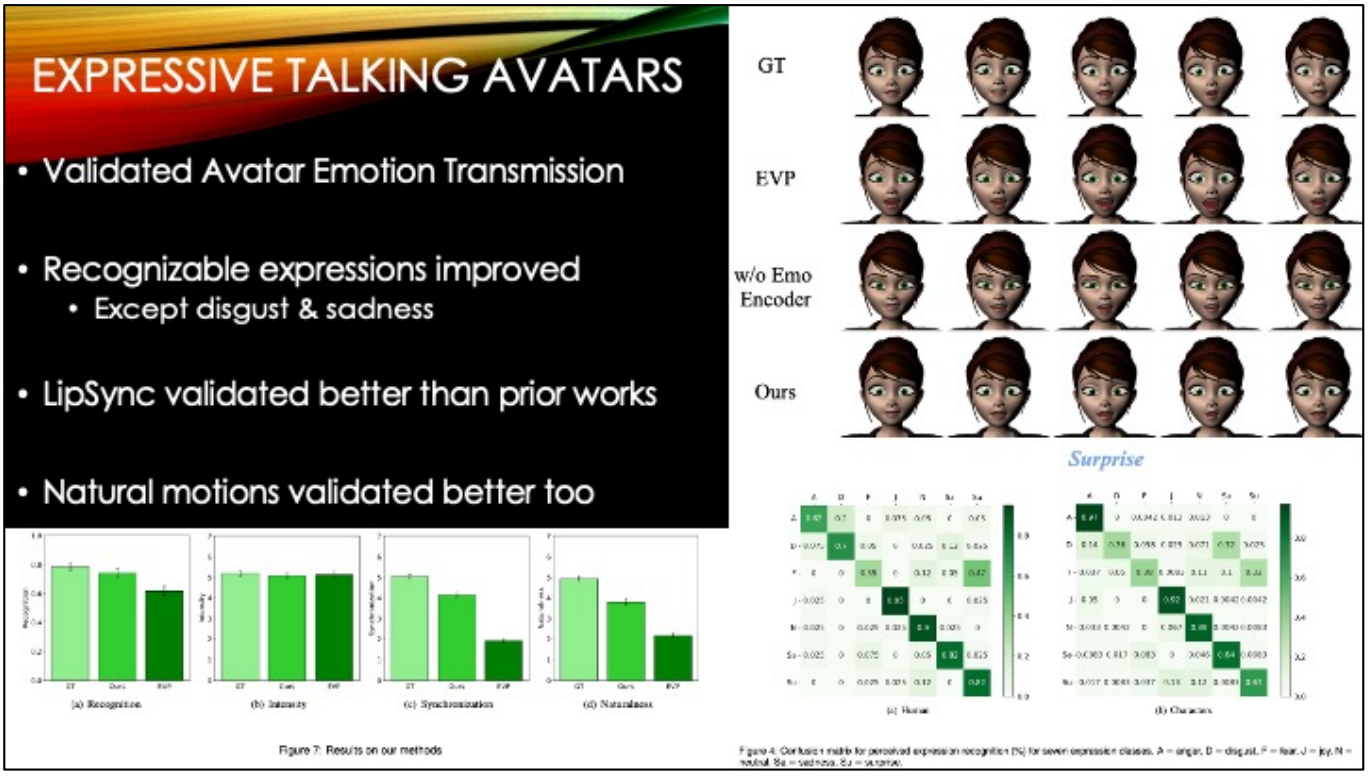
And to take away the main point,
this example showed that instead of typing text prompts in a web browser or tool for generative AI responses

we can already use speech recognition
and feed the recognized text tokens simultaneously
to both the responsive AI agent's voice
and drive the animation of a visual embodiment of that personna live in conversation
to have a more natural creative interaction

Of course, this language token sequence only context model is prone to some error,
like mistaking the nautical word 'sail' for a shopping kind of 'sale'
such allophones are hard to distinguish without deeper context,
but already we have quite a powerful way to interact naturally with generative AI.



But we can go further on our speech interactions for example,
what if the AI agent can tell how we're feeling and respond appropriately?
And at IEEE VR in 2024 we demonstrated just that in our MoodFlow work,
where our method employs sentiment analysis on the recognized text of the discussion,
to go along with the AI chat responses and provides contextual face
and body visual animations according the mood of the conversation
>click
So by fusing multi-modal inference methods in a synchronized system,
we bring interactions with potentially more empathy and substance to our immersive live applications.
And this video in the lower right is MoodFlow running on an Apple Vision Pro



And in any scientific works we need to validate our hypotheses

and we performed two user validation studies on how successful the animation of generative emotional expressions can be

To explain more, this further work on expressive talking avatars with facial animations generated from speech and emotion states.

With a group of 25 people we found an improved correlation between perceived expression recognition and the intended generated emotion over our prior emotional voice puppetry work with less focused data preparation,

Notable emotions here were disgust and sadness

being harder to perceive and distinguish as reliably in this experiment.

With our efforts on lip sync measures built into the model we developed

it was satisfying to see the statistically significant improvement from the prior work on the perceptual validation animation synchronized to speech audio.

And we also successfully improved upon the perceived naturalness of our animations

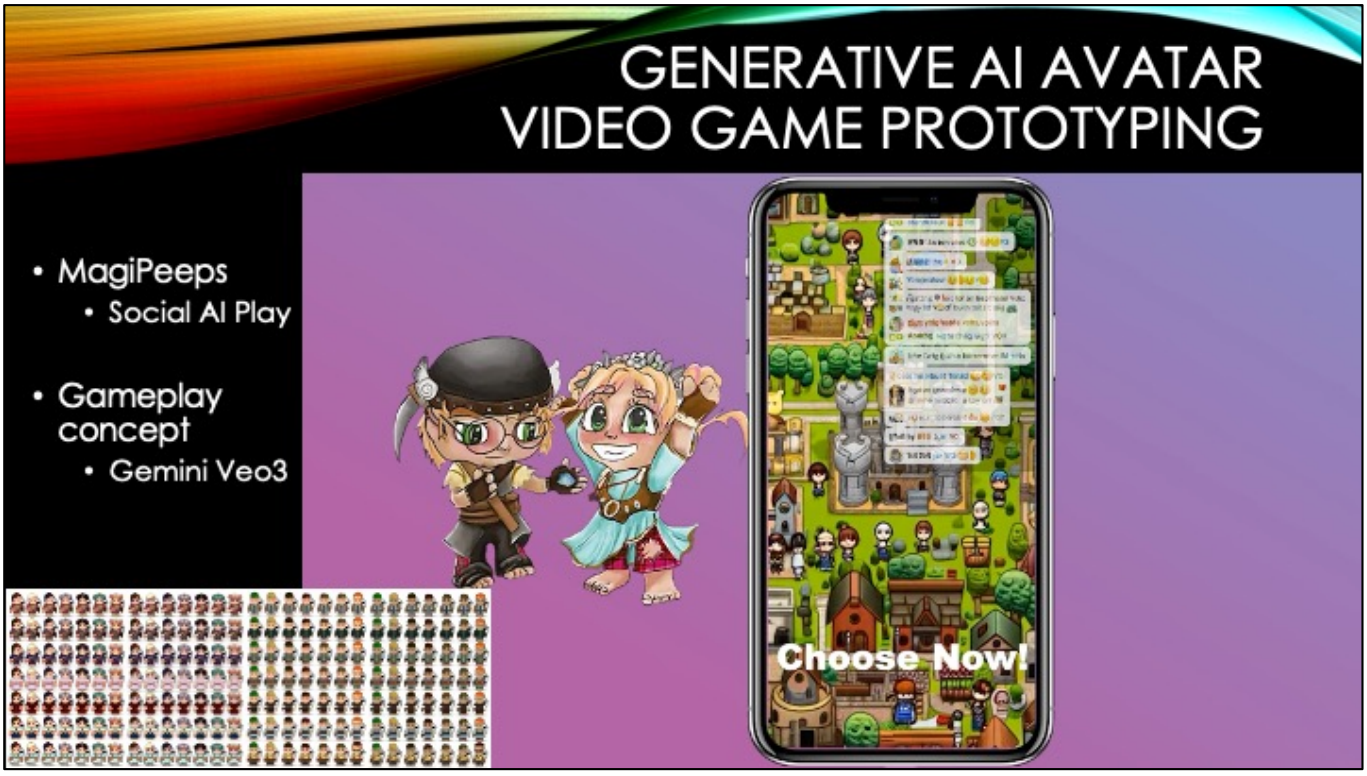
by developing our training loss function to be more continuity-oriented

To a great degree the source animation data content,

such as the FERG dataset used in this work,

hampers the work where the source expression data is really not the highest quality,

and this is an exciting potential of world leading artist's quality works from the past, present and future

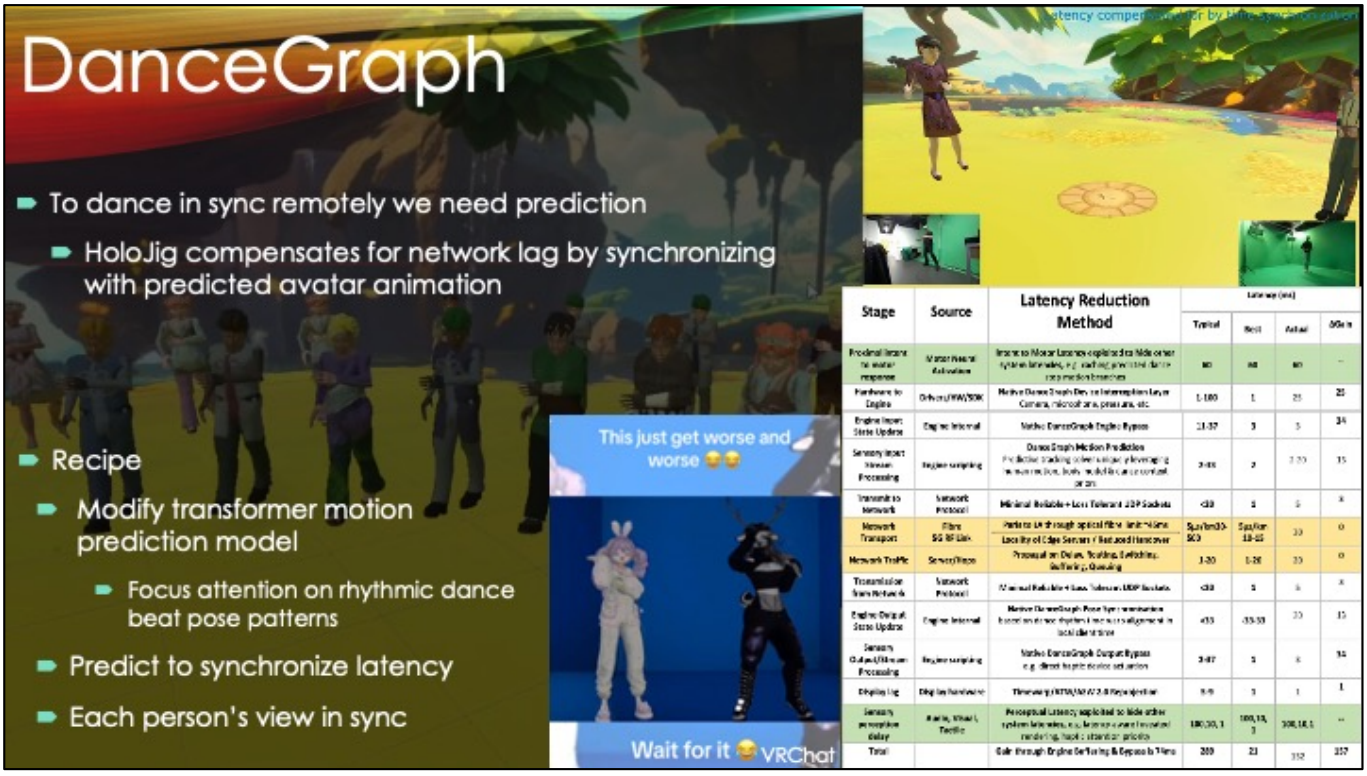


And staying on the theme of conversational characters and their social interactions, I'd like to show part of this pitch video experiment for a social sim genre video game.
as an attempt to bring together some of these generative ai approaches into video game development.

Fully AI generative to produce a concept video of a town gameplay design with our conversational avatars,
including the babble language stream text which I separately prompted and layered later in adobe premier.

And further using an agentic system in this case, gambo.ai
just as the simplest game design exercise,
and although it's very very far from usable for production.
it did give me some good tangible feedback on designing a game for network social interaction,
generating all the logos, sprite animations, music and sound effects and procedural maps
by running various agents to build assets and write integrated web code modules.
On the flip side it introduced bugs that it fixed in earlier chat iterations,
and the code wasn't designed well for scaling between desktop and mobile,
so I didn't go further beyond bouncing this simple idea off an AI generator

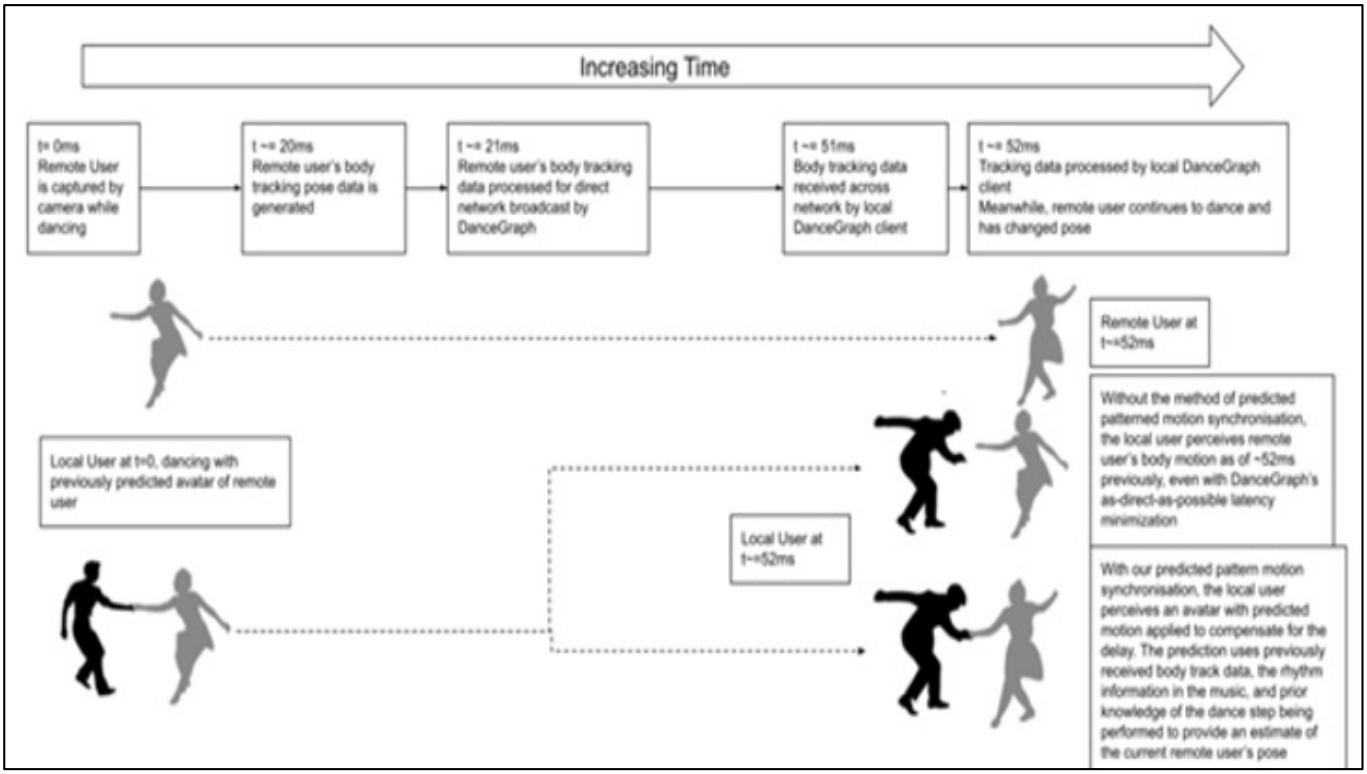
But also building further on this live conversational AI ability, why not try to go full Star Trek and build a holodeck (in VR)?
Instead of laboriously generating a traditional 3D scene using modeling tools and a team of artists with low quality results <here> in the upper middle,
we used stable diffusion to generate images from a prompt describing the world's concept, to produce a panorama image, which is a 360 view of a VR scene around you,
together with a depth map from this image generated automatically with the DepthAnything model
to provide all we need for a low-cost visual VR environment for dancing and other fun to take place in.



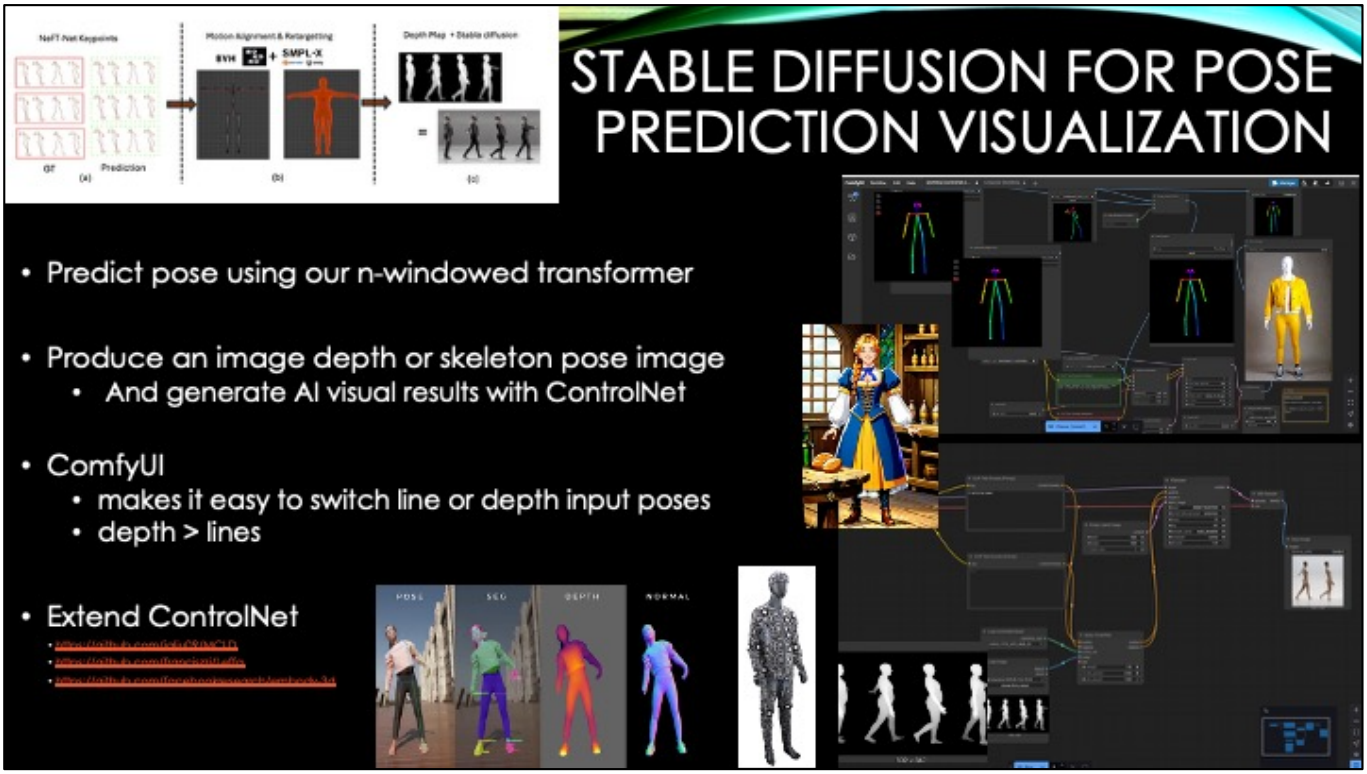
And in this dance research project the key breakthrough was on the live dance animation between friends
to be able to synchronize their moves even though the two partners are remotely located online in different countries
with all the latency and system lag you can see on the right
and commonly this shows up as a problem if you want to dance in sync in VR,
as shown with this TikTok video of VRChat dancers in the middle.

So, without too much detail, my solution here was to compensate for the latency between the dancers
by modifying a transformer model to focus attention on the rhythmic patterns inherent in the music's beat
allowing better animation pose predictions to then
use those predictions to synchronize each other together with the music in their VR headsets

Of course, the dancers are amateurs having fun and as a research prototype is it not perfect,
but the ability to be connect by being in step with each other thousands of miles apart is a solution to a basic physical speed of light problem inherent in any online interaction
and I'm quite happy to have achieved what we set out to do here.



And just to reiterate the problem solved again,
we illustrate the online dance prediction challenge along the timeline you see here.
Starting at t0 the local dancer is already dancing in sync with the remote predicted avatar and we're done.
…well not quite, the remote dancer at the same time is only just started capturing her pose and it's not until 20ms milliseconds or so have already elapsed only
the remote dancers' body tracking pose is just generated in the RAM of the remote computer,
as directly as possible we immediately present that pose data for network broadcast.
In good conditions and efficient network UDP socket code 30ms or so later we receive that remote dancer's pose on the local dancer's client PC
again at around 52ms in we have the remote dancer's pose ready to see
So, notice that local dancers sees his partners' pose as formed 50ms ago.
And the point here is critically
no matter what architectural measures we employ to reduce latency, we still have a lag however small,
and we still need to apply a motion prediction in order to 'keep up' with the local dancer's expected pose of his remote partner.
So, only with prediction using a combination of previously received tracked motions,
the knowledge of the dance music's beat and where the local dancer would expect his partner to be,
can we form a synchronous real-time dance in time with the music.



And ineed, we went further in visual terms to allow the possibility of posing realistic characters with generative AI,

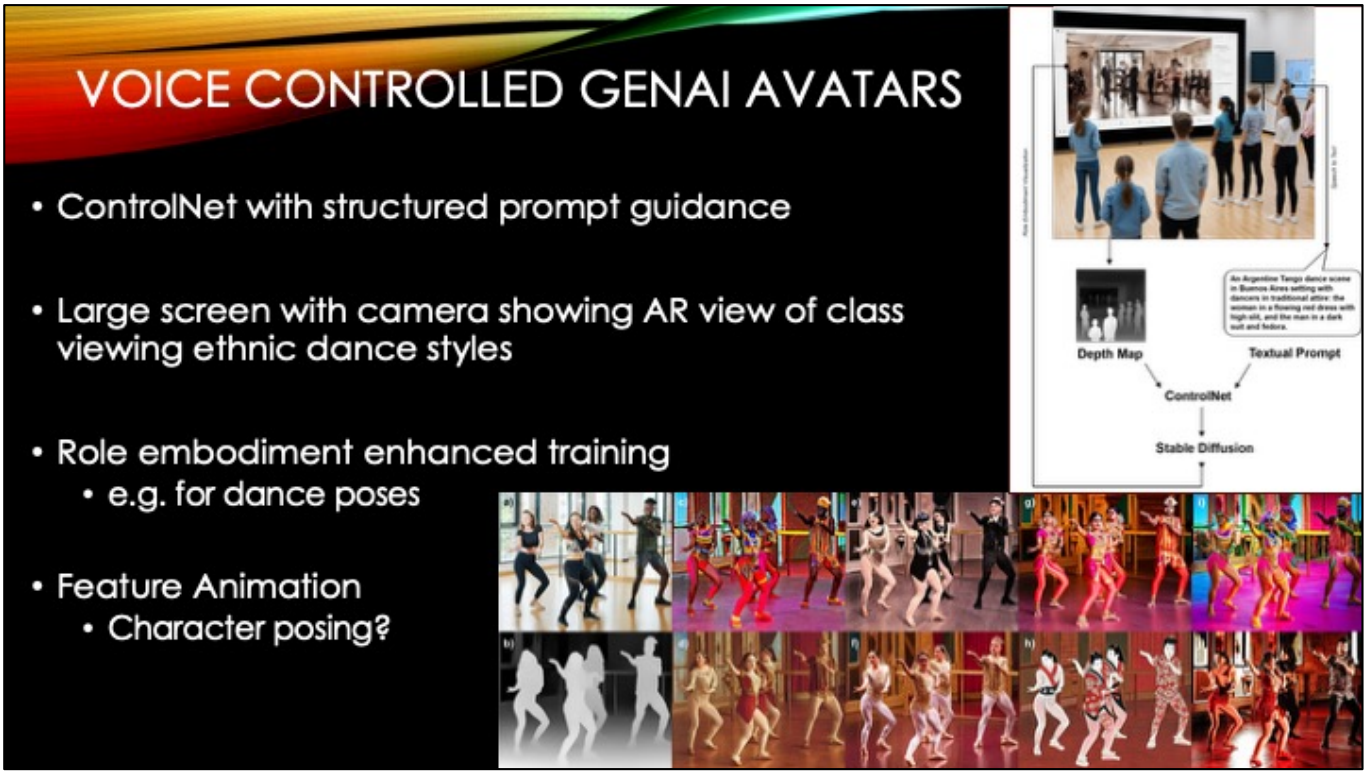
Here we provide our motion synchronized pose predictions mapped to either depth or skeleton pose images
and fed these to ControlNet to generate high-quality images of the avatars according to a prompted style.

It's worth noting we generated all frames of an animation sequence together to maintain consistency of results across all frames,
And I used the comfyUI framework to easily switch between depth control images or skeleton bone images and try out various models with different parameter controls.

We actually found with the line drawn bone images, the results were not that controlable
for getting the right ordering or orientation of the dancer's limbs
even with our direct bone length and orientation data
and the depth images where actually much better suited to these cases.

It's still interesting to consider building new ControlNet variations
and combining input signals to provide as much animation surface form information to the controlnet's conditioning input

I don’t think I’ve found a surface texture approach for animation control of the stable diffusion pose results,
but MCLD for example which uses a surface normal map signal, performs better than depth or skeleton posing already
I think it’d be interesting to look at motion capture suit patterns
for a detail orientated visual control signal and colorize with depth and orientation
MCLD is available on huggingface would be interesting to setup a comfyui node for that model and others
The embody dataset also released by Meta, NVIDIA just released a dataset, and these look like a great resource for detailed human body motion learning research



Moving to another practical use case, we applied this depth image oriented generative ai method in the classroom,

For example, here we integrated our speech genAI recipe with ControlNet and presented support for how to structure prompts for teaching uses cases for video mirror displays in the classroom.

Instead though here, we generated an inferred depth from a camera image in the classroom
and provided this depth along with a prompt to controlnet to yield a GenAI visualization of the students
as they would be dressed in a traditionally established dance style.

Although image updates here take 10 seconds with this current scheme it does provide a depiction of a specific pose that the teacher has set for the students to match,
which is often the practice in regular dance class training by example.

And this embodied learning style has been found to be much more effective than traditional classroom learning in related scientific studies

Of course, here it's also interesting to consider this kind of set up for developing concepts of characters

and how they might interact and pose in steps of a story board.

And as our approaches taking people into account is image based why not apply on the large scale with whole audiences?

Again this takes 10seconds or so,
but there's nothing to suppose we can't only generate key frames
and interpolation on motion predictions particularly for human dance,
to warp the frames for real-time animation.

Plus with depth and segmentation information, this informed live motion inbetweening can work really well.
And as real-time frame models generate next

A back to our vision of a generative AI metaverse in VR , perhaps whilst immersed in a holodeck VR space we can maintain a consistency of our experience whilst it transforms around us.

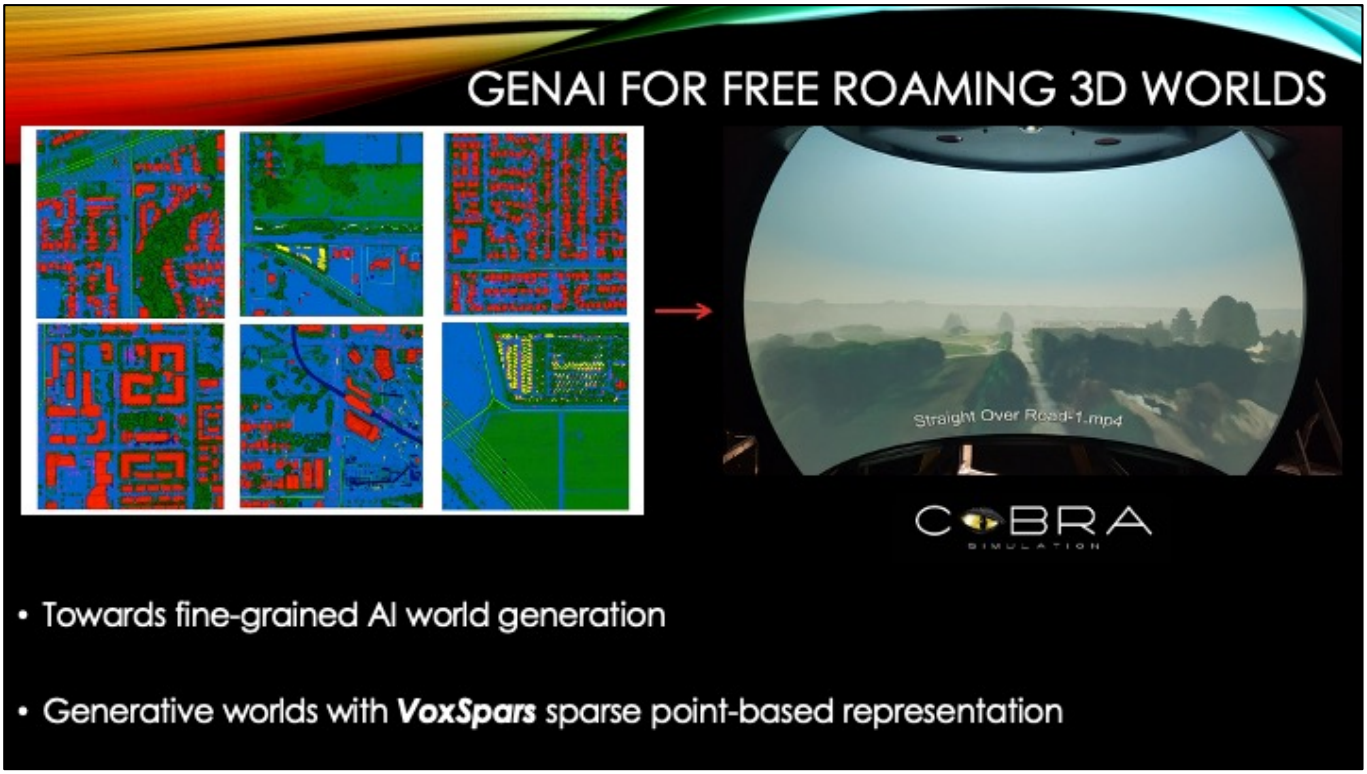
And so this is exactly what we've done this year, and super focused on the user experience of low latency
We applied measures to smooth the transition time whilst our genAI recipe in the background is computing.

Let's watch the video

And here's a better view of the depth parallax we generate for 3D views within the holodeck style space.

Very basically in detail here we extract low resolution intermediate stages of the gen AI image generation process
and apply it to a 3d effect in the VR space incrementally in real-time without halting the interaction, until the full resolution depth and image data is assembled.

And the bottom left shows these intermediate stages for various environments we prompted for, and their associated resolved depth generations



So wow we have conversational generative ai avatars that we talk and dance with and they listen to us with empathy, and we can jump into worlds that transform around us in real-time according to where we tell the computer to take us.

But that's of course the very beginning, things are still rough,
and we want more control for example what if we can edit an object in these worlds on demand
and at a fine-grained level of detail of our choosing.

This work here is for a company I CTO part time for, called Cobra Simulation in Scotland.
And we've developed this new technique for our Cobra World production called voxspars.
It's a little like the currently popular Gaussian splatting or nerf approaches you might have heard of, but much more tuned to real-time rendering with control in video game engines such as this example in unreal engine.

To explain, we take a details satellite scan of a 2 km region in Canada and instead of directly viewing as a point cloud that's quite sparse and low detail,
And we generate the inferred detail in real-time directly computing within each GPU shader fragment to fill in the gaps between points that are commonly one meter apart to produce a walkable, driveable version of the terrain right into the VR application.

In our case, that’s into the immersive dome display our company manufactures for a training drive-through of the generated photo realistic environment

And of course each point of the terrain can be modified in real-time to reflect changes in the environment for example if an explosion leaves a crator or we want to build a shed in the field or play golf.

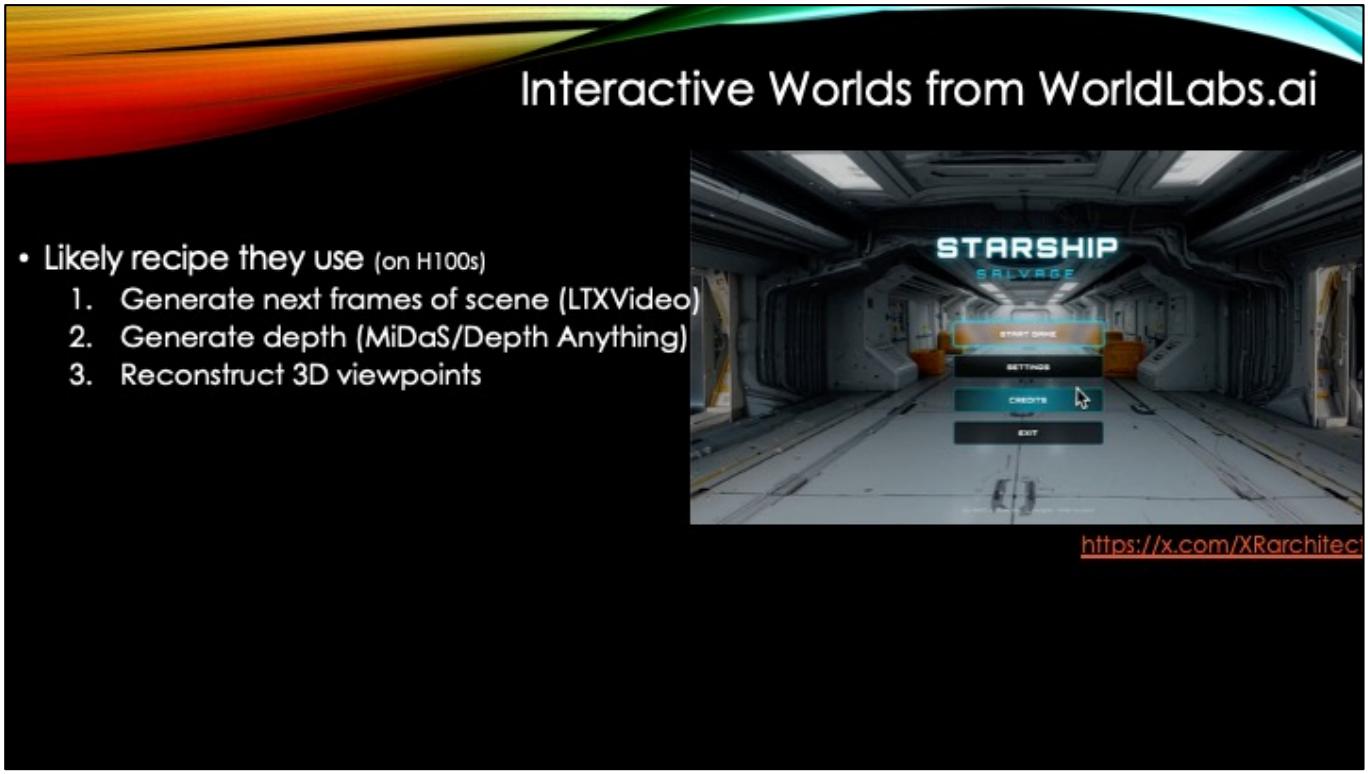
And coming right up to present, on this approach of providing sparse real data and filling in the gaps algorithmically.
I'm excited by this example I made using google gemini's terrific Nano Banna image generation model.

There's a lot of awesome prompted driven images being created, but here's one novel case
Given raw data scanned directly from a drone's LiDAR camera, for example of this street shown above,
We can rendered those points directly but you can see the buildings are missing walls
(as the drone from above missed those areas without seeing all the angles of the houses).
And notice the post boxes in the foreground are almost indecipherable as to what they might actually be.
But with that image as a reference to the generative ai model and together with a prompt suggesting the context of the scene, I got back this image on the bottom.

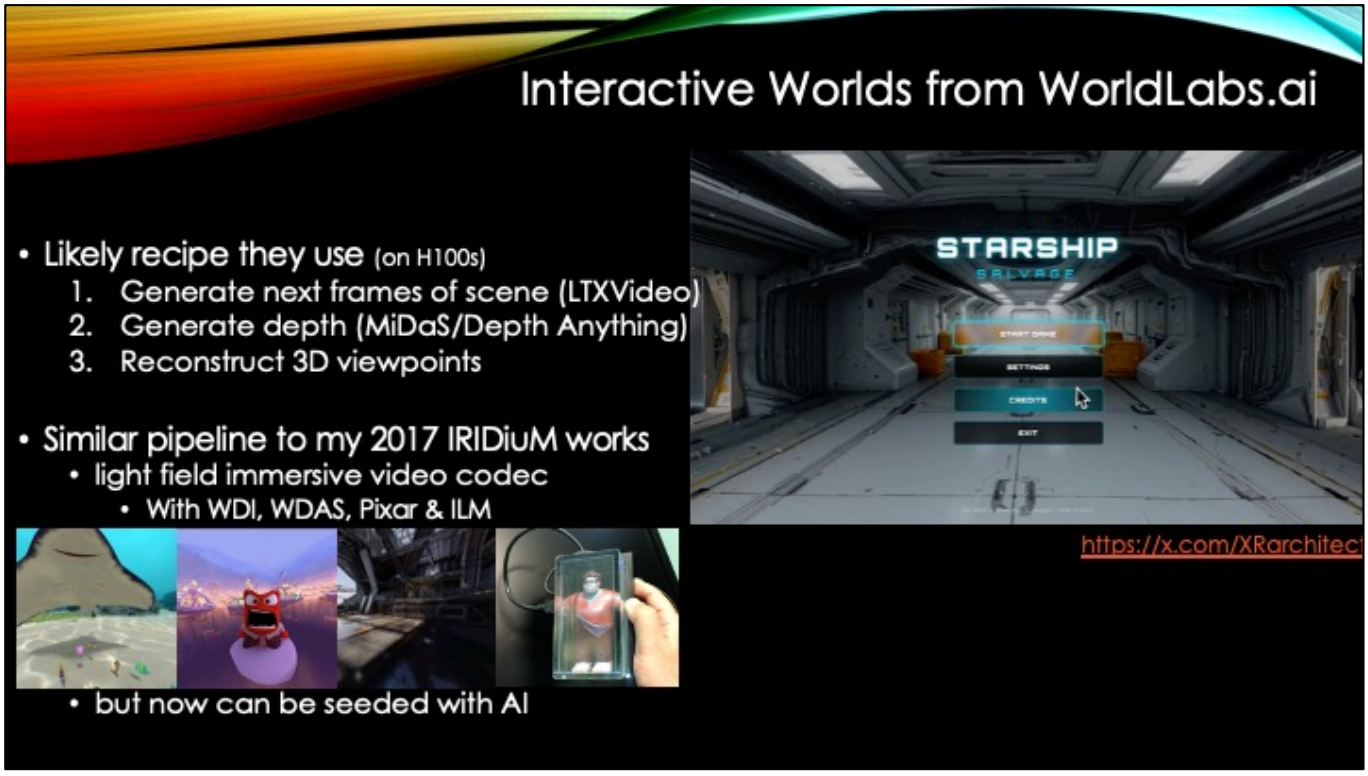
Not even needing to call out the missing wall it just filled in the gaps, and as you can see the post boxes looks pretty correct and with all the lighting generated photorealistically I must say this felt like an epiphany moment to me.
Even though it is hallucinating the result we can iterate in a supervised fashion and try out generations until we're convinced the results were really there as close to

authenticity to the original location as we need, a
nd of course we can apply creativity to change the appearance more
and steer the result for a video game environment generation or music videos, and so on.

Although again this is another example of the bitter lesson, where a whole research topic from computer vision majorly disrupted by these advances. For my voxspars methods it’s clearly a next step for me to apply these models for live reconstruction directly in the GPU fragment shader.



In this direction, I mentioned the Real Time Frame Model earlier ,
which is an approach in development by WorldLabs.ai that uses generated frames to produce next frames with spatial understanding.
And this is the recipe I believe they're using to produce results like those on the right.
Using video diffusion transformers to generate next frames of animation,
Making depth from those images,
and then reconstructing 3D view points combine those depths with color mapped in 3D.



Which is actually almost identical to my IRIDiuM immersive light field work for Disney Research research about 8 years ago,
That built animated 3d views from depth, color and reflectance data
and went really deep into optimizing for video codec efficiency and adaptively breaking down into tiles for memory efficiency.
Just now we can seed these interactive animation generations AI content

And of course there's others developing this topic right now, including Tencent with their flashworlds which generate 3D guassian splat scenes in 7 seconds on A100 gpus.

So that's hopefully given you all a flavor of some recent results towards making generative ai more interactive and that's there so much more to do in the exciting growing area to transform how we make content across all of games, movies and beyond