MY WEIRD PROMPTS

Podcast Transcript

EPISODE #132

Beyond Frames: The Rise of Real-Time Video AI

Published January 02, 2026 • Runtime: 22:06

https://myweirdprompts.com/episode/video-multimodal-ai-evolution/

EPISODE SYNOPSIS

In this episode of My Weird Prompts, hosts Herman and Corn dive into the cutting-edge landscape of 2026's video-based multimodal AI. They explore how the industry moved beyond simple frame-sampling to adopt spatial-temporal tokenization, allowing models to treat time as a physical dimension. The discussion covers the technical hurdles of real-time video-to-video interaction, including Simultaneous Localization and Mapping (SLAM) for floor plan generation and the use of speculative decoding to minimize latency. By examining the integration of Neural Radiance Fields (NeRFs) and native multimodality, Herman and Corn reveal how AI is finally crossing the uncanny valley to create digital avatars that are indistinguishable from reality.

DANIEL'S PROMPT

Daniel

I'd love to hear your thoughts on the advent of video-based multimodal AI. Video consists of a series of images, requiring a huge amount of data and computation for a model to process alongside a prompt. I've been thinking about workflows like using Gemini 3 to generate a floor plan by walking through an apartment, which involves depth mapping and complex video processing. Perhaps the most challenging use case is real-time video-to-video AI, such as interacting with a high-fidelity, indistinguishable avatar. How is this achieved from a context and tokenization standpoint, and what can we look forward to as these models become more realistic?

TRANSCRIPT

Corn

Hey everyone, welcome back to My Weird Prompts! I am Corn, and I am sitting here in our living room in Jerusalem with my brother.

Herman

Herman Poppleberry, at your service. It is a beautiful day to dive into some high-level technical discourse.

Corn

It really is. And our housemate Daniel sent us a fascinating audio prompt this morning. He was walking around the apartment trying to use Gemini three to generate a floor plan, and it got him thinking about the sheer complexity of video-based multimodal models.

Herman

I love that Daniel is basically stress testing the state of the art just by walking to the kitchen. But he is right to be curious. We are living in two thousand twenty-six, and the leap we have seen in video processing over the last twelve months is nothing short of miraculous. When you think about where we were just a couple of years ago, it is staggering.

Corn

Exactly. Daniel was asking about the advent of video-based multimodal artificial intelligence. He pointed out that video is really just a series of images, which means a huge amount of data. He wanted to know about the tokenization, the context windows, and specifically that holy grail of real-time video-to-video interaction, like talking to an indistinguishable avatar.

Herman

That is the deep end of the pool, Corn. I have been looking at some of the recent papers on spatial-temporal tokenization, and it is a complete rethink of how we handle information. Most people think of video as just a stack of photos, but for a model to actually understand it, it has to treat time as a dimension just as important as height or width.

Corn

Right, and that is where I want to start. If we are talking about twenty-four frames per second, or even sixty frames per second, a one-minute video is thousands of images. If each image is a set of tokens, how does the model not just explode under the weight of all that data?

Herman

That is the big question. In the early days, we used to just sample frames. We would take one frame every second and hope the model could stitch the context together. But today, with models like Gemini three and the latest versions of the Claude and GPT families, we are using something called Vision Transformers with three-dimensional patches.

Corn

Explain that a bit more for me. When you say three-dimensional patches, you are not just talking about the image itself, right?

Herman

Exactly. Imagine a cube of data. Instead of just taking a square patch of an image, like a sixteen by sixteen pixel block, the model takes a volume. It takes that sixteen by sixteen square and extends it through time, say, across eight or sixteen frames. So, one token actually represents a small window of both space and time. This is what we call temporal compression. It allows the model to see motion within a single token rather than having to compare two separate tokens to figure out what moved.

Corn

That is fascinating. So, the token itself contains the information that something is sliding or rotating?

Herman

Precisely. It captures the essence of the change. This drastically reduces the total token count. If we tried to tokenize every single pixel in every single frame individually, even the massive context windows we have now, which are reaching into the tens of millions of tokens, would be filled up in minutes.

Corn

Okay, so that handles the efficiency part of the equation. But Daniel's example was specifically about walking through the apartment to create a floor plan. That feels like it requires a different kind of understanding. It is not just seeing motion, it is building a mental map of a three-dimensional space from a two-dimensional video feed.

Herman

You are hitting on the difference between video recognition and world modeling. When Daniel walks through the house, the model has to perform what we call Simultaneous Localization and Mapping, or SLAM, but it is doing it through the lens of a Large Language Model. It is identifying the corners of the room, the depth of the hallway, and the scale of the furniture. It is using those spatial-temporal tokens to realize that as the camera moves left, the objects on the right should disappear in a consistent way.

Corn

It is almost like the model is hallucinating a three-dimensional structure and then checking it against the video frames as they come in.

Herman

That is a great way to put it. It is maintaining a latent representation of the room. When Daniel turns a corner, the model is not just seeing a new image, it is updating its internal map. This is why the context window is so vital. If the model forgets what the front door looked like by the time Daniel reaches the balcony, the floor plan falls apart.

Corn

And we are seeing context windows now that can hold hours of video data. I remember reading that some of these experimental models can process up to ten million tokens. That is enough to hold the entire history of a long walk through a building with high fidelity.

Herman

It really is. But the compute required is still immense. We are talking about clusters of specialized chips working in parallel just to keep that context active. And this brings us to the second part of Daniel’s prompt, which is the real-time aspect. Doing this on a recorded video is one thing. Doing it live, where the AI has to respond to you as you are moving, is a whole different beast.

Corn

Yeah, the latency issue. If I am talking to a high-fidelity avatar, and there is a two-second delay while the model tokenizes my face and decides how to react, the illusion is broken immediately.

Herman

Exactly. The uncanny valley is not just about how the avatar looks, it is about how it feels in time. If the rhythm of the conversation is off, your brain flags it as fake.

Corn

Let us hold that thought on the real-time avatars, because I want to dig into the hardware and the inference speed that makes that possible. But first, we need to take a quick break for our sponsors. Larry: Are you worried about the government, or perhaps your neighbors, using advanced lidar to map your living room while you sleep? Do you feel like your personal space is being invaded by invisible data streams? Well, worry no more! Introducing the Quantum Privacy Umbrella. This is not just an umbrella, it is a localized signal-refraction shield. Simply deploy the Quantum Privacy Umbrella in the center of your room, and its patented micro-mesh of lead-infused silk will scatter all incoming and outgoing scanning frequencies. It creates a literal blind spot in the digital fabric of reality. Perfect for high-stakes meetings, naps, or just hiding from the future. It even comes in a stylish matte black finish that looks great in any bunker. The Quantum Privacy Umbrella, because if they cannot see you, they cannot tokenize you! BUY NOW!

Corn

Alright, thanks Larry. I am not sure lead-infused silk is actually a thing, but I guess that is why he is the ad guy.

Herman

I would be more worried about the weight of that umbrella than the lidar scans, to be honest. Anyway, back to the real-time video-to-video challenge.

Corn

Right. So, Daniel was talking about interacting with an indistinguishable avatar. For that to happen, the model has to do three things simultaneously: it has to ingest your video, understand your intent and emotion, and then generate a high-fidelity video response, all in under a hundred milliseconds.

Herman

That hundred-millisecond threshold is the gold standard. That is roughly the limit of human perception for what feels like real-time. To achieve this, we have moved away from the traditional pipeline where the model finishes one task before starting the next.

Corn

You mean it is not just a chain of command anymore?

Herman

No, it is more like a fluid stream. We use something called speculative decoding and streaming inference. As the model is receiving the first few frames of your movement or the first few syllables of your speech, it is already starting to predict the likely end of that movement or sentence. It begins generating the response video before you have even finished your thought.

Corn

That sounds incredibly risky. What if the user changes their mind or does something unexpected?

Herman

Then the model has to pivot instantly. It throws away the predicted frames and re-calculates. This is why the compute demand is so high. You are essentially running multiple versions of the future in parallel and showing the user the one that matches reality.

Corn

And the avatar itself? Daniel mentioned it being indistinguishable from a person. That requires more than just good textures. It requires micro-expressions, the way light bounces off the skin, the way eyes reflect the environment.

Herman

That is where the generative part of the multimodal model shines. In twenty-six, we are seeing the integration of neural radiance fields, or NeRFs, directly into the generative process. Instead of just drawing a flat image, the model is essentially rendering a three-dimensional volume in real-time. This allows the avatar to have consistent lighting even if the user moves their camera around.

Corn

So, if I move my phone while talking to an AI avatar, the shadows on the avatar’s face will shift correctly because the AI knows where the virtual light source is in relation to my phone’s position?

Herman

Exactly. It creates a sense of shared physical space. That is what makes it indistinguishable. It is not just a video playing back at you, it is a dynamic entity reacting to the physics of your world.

Corn

I find it interesting that Daniel brought up the floor plan example alongside the avatar example. On the surface, they seem different, but they are both about understanding space and time. One is about mapping the world, the other is about existing within it.

Herman

They are two sides of the same coin. Both require the model to have a deep, intuitive understanding of three-dimensional geometry. When the model generates a floor plan from a video, it is proving it understands the constraints of physical reality. It knows that a wall cannot just end in mid-air and that a doorway has to lead somewhere. That same understanding is what allows it to generate a realistic human face that doesn't warp or glitch when it turns to the side.

Corn

What about the tokenization of the audio in this mix? Daniel’s prompt was an audio file. In a real-time video-to-video interaction, the audio and video have to be perfectly synced.

Herman

That is the beauty of true multimodality. In older systems, you had a vision model, a speech-to-text model, a language model, and a text-to-speech model all taped together. There were huge delays at every junction. Today, we have native multimodal models. The audio, the video, and the text are all converted into the same token space from the very beginning.

Corn

So the model "sees" the sound waves and the pixels as the same kind of information?

Herman

In a sense, yes. It is all just patterns in a high-dimensional space. This allows for what we call cross-modal attention. The model can use the visual cue of your lips moving to help it disambiguate a muffled sound in the audio. It is much more robust, just like how humans use their eyes to help them hear in a noisy room.

Corn

That explains why the latest avatars are so much better at lip-syncing. They aren't just trying to match a sound to a mouth shape, they are generating the face and the voice as a single, unified output.

Herman

Exactly. And the implications for this are huge. Think about telepresence. Instead of a flat Zoom call, you could have a real-time, three-dimensional avatar of your colleague sitting in the chair across from you, rendered with perfect fidelity, reacting to your movements and the lighting in your room.

Corn

It makes me wonder about the data side of things. Daniel mentioned the huge amount of data and computation. Where does all this training data come from for video? We have already scraped most of the text on the internet. Is the next frontier just every YouTube video and every security camera feed?

Herman

Pretty much. But it is not just about quantity anymore, it is about quality. We are seeing a lot of work in synthetic data. We use physics engines to generate millions of hours of perfectly labeled video. If you want a model to understand how a glass breaks, you can simulate it a thousand different ways in a game engine and feed that to the model. That gives it a grounded understanding of physics that it might not get from just watching random home movies.

Corn

So the model learns the laws of gravity and momentum through these simulations?

Herman

Yes. We call this the world model approach. By training on video, the AI is essentially learning a simplified version of physics. It knows that if you drop a ball, it goes down. It knows that if you walk behind a pillar, you should reappear on the other side. This is why Daniel's floor plan experiment works. The model has an internal sense of how space is structured.

Corn

It is a bit mind-blowing when you think about it. We started with chatbots that could barely remember the previous sentence, and now we are talking about entities that understand the three-dimensional layout of our homes and can simulate human presence in real-time.

Herman

It is a massive shift. And we are just at the beginning of the video-to-video era. By twenty-seven, I expect we will see these models running locally on high-end hardware. You won't even need a massive server farm to talk to a basic avatar.

Corn

That brings up a good point about practical takeaways. If I am a listener and I am hearing all this, what does it mean for me today, in early twenty-six?

Herman

Well, for one, it means your workflows are about to become much more visual. If you are a designer, a contractor, or even just someone moving house like Daniel, you won't be drawing things by hand as much. You will be showing the AI your world, and it will be helping you manipulate it.

Corn

I think the biggest takeaway for me is the shift in how we interact with information. We are moving away from the "search and click" model toward a "show and tell" model. Instead of typing "how do I fix this leak," you will just point your camera at the pipe, and an avatar will appear to walk you through the repair in real-time, pointing at exactly where you need to put the wrench.

Herman

That is the killer app right there. Expert knowledge, delivered through a human-like interface, with full spatial awareness. It democratizes skills that used to take years to master.

Corn

But we should probably talk about the risks, too. If the avatars are indistinguishable, the potential for deepfakes and misinformation is off the charts. We are already seeing some pretty sophisticated scams using real-time video.

Herman

You are absolutely right. The technology to detect these models has to evolve as fast as the models themselves. We are looking at digital watermarking at the sensor level, where your camera signs the video feed to prove it is real. But even then, it is going to be a constant arms race.

Corn

It is a strange world we are building. I think about Daniel walking through the apartment, and I wonder if he realizes that he is essentially teaching the AI how to live in our house.

Herman

He probably does. Daniel is always three steps ahead. But that is the nature of this collaboration. We provide the data of our lives, and the AI provides us with new ways to understand and navigate that data.

Corn

Well, I think we have covered a lot of ground here. From three-dimensional token patches to real-time physics simulations and lead-infused umbrellas.

Herman

Don't forget the hundred-millisecond latency barrier. That is the one to watch. Once we consistently break that, the line between the digital and the physical is going to get very, very blurry.

Corn

It is already getting blurry for me. I am starting to wonder if you are a real-time avatar, Herman. Your lip-syncing has been suspiciously good today.

Herman

If I were an avatar, would I know? That is a question for another episode. But I can assure you, my nerdiness is one hundred percent organic.

Corn

Fair enough. Before we wrap up, let us give some practical advice for people who want to play with these models. If you are using something like Gemini three or the latest multimodal tools, try to give them as much spatial context as possible. Move the camera slowly, show the depth of the room, and use descriptive language. The more you help the model build its internal map, the better your results will be.

Herman

And stay curious. The field is moving so fast that what we said today might be old news in six months. Keep an eye on the research coming out of the big labs regarding temporal consistency. That is the next big hurdle.

Corn

Absolutely. Well, I think that is a wrap for today. A huge thank you to Daniel for sending in such a provocative prompt. It really forced us to dig into the guts of how these video models work.

Herman

It was a blast. I could talk about spatial-temporal tokens all day, but I think the listeners might appreciate a break.

Corn

Thanks for listening to My Weird Prompts. If you enjoyed this deep dive, you can find more episodes on Spotify or visit our website at myweirdprompts.com. We have an RSS feed there for subscribers, and a contact form if you want to send us your own weird prompts. We would love to hear what you are thinking about.

Herman

Until next time, I am Herman Poppleberry.

Corn

And I am Corn. We will see you in the next episode. Stay curious, everyone!

Herman

And keep your lidar-blocking umbrellas handy!

Corn

Goodbye! Larry: BUY NOW!