Episode Book

EPISODE #5625

Omni and the Marble That Rolls Uphill

PUBLISHED	28 September 2026 at 14:25 UTC
DURATION	26:11
LISTEN	myweirdprompts.com/episode/omni-physics-world-models/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	deepseek-flash
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.2
BOOK GENERATED	19 October 2026 at 17:38 UTC

SYNOPSIS

Google DeepMind's Gemini Omni landing page claims the model has "an intuitive understanding of forces like gravity, kinetic energy, and fluid dynamics." But video models have been walking people through walls for years — and a landmark ICML 2025 paper found they generalize by case-based retrieval, not physical law. So how did Omni seem to close the gap without the radical architectural departure the world models camp, including Yann LeCun, insisted was necessary? This episode unpacks the claim, the failure history, the counter-argument, and what Omni actually did differently.

TRANSCRIPT



Corn

Okay, so the thing about the marble is that it has to slow down on the incline, and every time it doesn't, it looks wrong immediately.



Herman

Right, and that's the whole problem in miniature, isn't it? You can render a marble perfectly. You can get the light bouncing off it, the reflection of the room, the way the wood grain blurs underneath it. Every pixel can be flawless.



Corn

And then it rolls uphill.



Herman

Which is exactly the thing Daniel's asking about. He sent us a prompt this week about Google's Omni, and the claim on the DeepMind landing page that it has, quote, "an intuitive understanding of forces like gravity, kinetic energy, and fluid dynamics for more realistic movement."



Corn

He's got four things going at once in there, which is very him. First, what does that claim actually mean. Second, why it matters, given that frontier models have been generating people walking through walls for years. Third, the skeptical argument that large language models can't understand the world as a physical construct because they only see patterns in knowledge, and that you need something called world models to get past that. And fourth, the one he's really circling, which is how Omni seems to have addressed all of that without the radical architectural departure the world models camp said was necessary.



Herman

That last one is the actual question, isn't it. The first three are context. The fourth one is the bet.



Corn

So today we're going to unpack the claim, the history of the failure, the counter-argument, and what Omni actually did differently. Let's start with what Omni is, because the wording of the claim matters as much as the claim itself.



Herman

Gemini Omni is DeepMind's flagship multimodal generative model. The tagline on the page is "Create anything from any input, starting with video." It's been described as Nano Banana but for video, which is a decent shorthand. The current public version is Gemini Omni Flash, and it's in the Gemini app, Google Flow, YouTube Shorts, Google Vids, and AI Studio.



Corn

And the physics claim is sitting right there on the landing page, in a section called "Bring ideas to life, grounded in Gemini's world knowledge." The verbatim line is the one Daniel quoted. "Create output that follows real-world physics. Omni has an intuitive understanding of forces like gravity, kinetic energy, and fluid dynamics for more realistic movement."



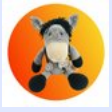
Herman

Now read the next section header on that same page. "Apply real world knowledge." And the sentence under it says Omni combines an intuitive understanding of physics with Gemini's knowledge of history, science, and cultural context, bridging the gap from photorealism to meaningful storytelling.



Corn

So physics is one facet of a unified world-knowledge representation. It's not a module they bolted on.



Herman

Right. And here's the tell. DeepMind describes Omni as delivering, quote, "a leap in world understanding, multimodality, and editing." World understanding. Not world model.



Corn

That word choice is the entire episode compressed into two syllables.



Herman

It is. Because the world models camp has a very specific thing in mind when they say world model, and it is not what DeepMind is claiming here. DeepMind is claiming a capability. The world models people are claiming you need an architecture.



Corn

Give me the concrete capabilities on the page, because the demos tell you what they think the model can do.



Herman

The headline demo is a marble rolling on a chain-reaction track. Continuous smooth shot. Then there's multi-turn editing with scene consistency. You can say make the violin invisible, or change the camera angle to be over the violinist's shoulder, and the scene holds together across turns. Reference-to-video, where you combine an image, an audio clip, and text into one output. First and last frame specification. Video extension. You can draft at 360p and upscale to 1080p or 4K. And every output carries SynthID watermarking and C2PA Content Credentials.



Corn

The marble on the chain-reaction track. That's not an accident. That's the demo they chose because it's the one where physics failure would be most obvious.



Herman

A chain reaction is a physics test. If the marble doesn't transfer momentum correctly, if it doesn't slow on the incline, if it clips through a piece of track, the whole thing falls apart visually. So they're putting their strongest physics example front and center.



Corn

Before we can evaluate whether Omni solved the physics problem, we need to understand exactly how badly models were failing at it, and why.



Herman

So the phenomenon has a name in the literature. Visual realism but physical absurdity. Generated video that looks photorealistic and violates basic mechanics. People walking through walls. Objects passing through each other. Water flowing uphill. Gravity that changes direction between cuts.



Corn

The walking through walls thing is the canonical example because it's so stark. You get a person with perfect skin texture, perfect hair, perfect clothing folds, and then they just phase through a load-bearing wall like it's fog.



Herman

And it's not a rendering error. The model rendered the wall beautifully. It rendered the person beautifully. It just didn't have any representation of the fact that two solid objects can't occupy the same space.



Corn

There's a paper from ICML 2025 that's the landmark study here. "How Far is Video Generation from World Model: A Physical Law Perspective." Kang, Yue, Lu and others.



Herman

The experimental design is what makes it useful. They built a 2D simulation testbed governed by classical mechanics, and they trained diffusion video models to predict object motion inside it. So it's a controlled environment where you know exactly what the physics should be, and you can watch the model either get it right or get it wrong.



Corn

And the results?



Herman

Three findings, and they escalate. Perfect generalization within the distribution. Measurable scaling behavior for combinatorial generalization. And then failure in out-of-distribution scenarios.



Corn

So if the test looks like the training data, the model nails it. If it's a novel combination of things it's seen, it does okay and gets better with scale. If it's outside what it's seen, it falls apart.



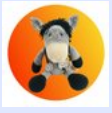
Herman

And the two insights underneath that are the important part. First, the models fail to abstract general physical rules. Instead they exhibit what the paper calls case-based generalization. Which means they're mimicking the closest training example rather than applying a law.



Corn

Say that again, because it's the whole ballgame.



Herman

Case-based generalization. The model isn't reasoning about momentum. It's retrieving the most similar clip it saw during training and reproducing that motion pattern. If the new situation is close enough to a training example, the output looks physical. If it isn't, the output looks like whatever the nearest neighbor was, which may be completely wrong.



Corn

So it's not simulating. It's remembering.



Herman

It's remembering, and the remembering is good enough to fool you most of the time. That's why the failure is so jarring when it happens. You've been watching outputs that look right for twenty minutes, and then something impossible happens, and you realize none of it was grounded.



Corn

What's the second insight?



Herman

When the models do generalize, they prioritize the wrong factors. The paper found the ordering is color, then size, then velocity, then shape.



Corn

Color first. The most physically irrelevant variable on the list.



Herman

Color first. The model latches onto superficial appearance cues over the variables that actually determine how something moves. Shape is last, which is almost comical, because shape is what determines collision behavior in most of these scenarios.



Corn

And the paper's bottom line?



Herman

Quote. "Scaling alone is insufficient for video generation models to uncover fundamental physical laws, despite its role in Sora's broader success."



Corn

That's a direct shot at the scaling thesis. More compute, more data, bigger model, and you still don't get physical law.



Herman

And this isn't one paper's quirk. There are supporting benchmarks that tell the same story. WorldBench isolates single physical concepts. Object permanence. Friction coefficients. Fluid viscosity. One concept at a time, so you can see exactly where the model breaks. And the finding across the board is that all tested models lack the physical consistency required to generate reliable real-world interactions.



Corn

All tested models.



Herman

Then there's PhyWorld, which frames the requirement as preserving the physical state implied by the conditioning input, and notes that this isn't reliably achieved yet. And GEM-4D, which found that video world models often fail to track the same physical points consistently across time. They appear plausible but lack the physical grounding required for reliable action execution.



Corn

The GEM-4D number that stuck with me is the robot one.



Herman

Adding geometry grounding improved real-world robot manipulation success from sixty-one percent to eighty-one percent. Twenty points from giving the model actual spatial structure instead of asking it to infer it from pixels.



Corn

So the mechanism explanation, because this is where it stops being a list of failures and starts being a diagnosis. Why does this happen?



Herman

Because the training objective is pixel prediction, not dynamics simulation. The model is rewarded for producing frames that look like the training distribution. It is not rewarded for conserving momentum. It is not rewarded for respecting collision boundaries. It is not rewarded for anything except making the next frame look plausible.



Corn

And case-based generalization is just the technical name for what that produces. What looks like physics but is really nearest-neighbor retrieval over visual patterns.



Herman

The model has learned an extraordinarily rich prior over what video looks like. It has not learned the rules that generate video. Those are different things, and the gap between them is exactly where the walking-through-walls failures live.



Corn

So that's the failure. Now here's the argument that says you can't fix it without throwing out the whole architecture, and here's why Omni suggests you might not have to.



Herman

The world models argument, stated cleanly. There's a paper, PAN, that puts it this way. World models represent the next frontier beyond large language models to enable physical and embodied intelligence. That's the thesis. LLMs got us language, and



Herman (continued)

language is a lossy, low-bandwidth projection of reality. Text describes the world. It doesn't contain the world. So if you want a system that can act in physical space, you need a different kind of model.



Corn

And the strongest articulation of that is LeCun.



Herman

Yann LeCun has been arguing for years that LLMs are a dead end for physical and embodied intelligence. His point is that they're trained on text, and text is a compressed description of reality written by humans for humans. It's missing almost everything about how the world actually evolves moment to moment. And in March he raised a billion dollars to build AI that understands the physical world.



Corn

A billion dollars is not a hedge. That's a conviction bet.



Herman

It's a bet that world models are a separate path from LLMs, not a continuation of them. His approach is JEPA, and V-JEPA for video. The idea is to learn abstract representations from video rather than predicting pixels or tokens. Don't try to generate the next frame. Learn the structure that generates frames.



Corn

So the world models camp is saying the architecture is wrong, not the scale. You can't get there from here.



Herman

That's the claim. And the reason it's a serious claim is that the failure mode we just described is exactly what you'd predict if the architecture were wrong. Case-based generalization is what you get when the model's only job is to make plausible-looking output.



Corn

Now give me the counterpoint, because this is where people usually stop.



Herman

V-JEPA 2, LeCun's own approach, has documented limitations. It's been described as basically a diva about camera positioning. It has long-horizon drift. It hallucinates beyond a few planning steps. So the radical departure path is not obviously winning either.



Herman

It is, because it means the model is sensitive to viewpoint in a way that a real physical understanding wouldn't be. If you actually understood the scene in three dimensions, camera position would be a rendering parameter, not a failure pattern.



Corn

And there's a more fundamental point about the debate itself.



Herman

There is, and it's the one that keeps the argument honest. Just because you failed to elicit a capability doesn't mean you proved it cannot be elicited. Capabilities in these models are jagged and prompt-dependent. You can find a task where a model fails completely and then find a rephrasing where it succeeds. So the absence of physics understanding in a given eval is evidence, but it's not proof of an architectural ceiling.



Corn

Which cuts both ways. It means the world models people can't declare victory from the failure cases, and it means the scaling people can't declare victory from the success cases.



Herman

Now here's the thesis for Omni, and I think it's the answer to Daniel's actual question. Omni's approach is physics grounding as an emergent property of large-scale multimodal training, not a separate architecture.



Corn

Three pieces of evidence.



Herman

First, no new architecture is claimed. Omni is built on the Gemini family. Same transformer-based multimodal stack. DeepMind's language is intuitive understanding and world understanding. They deliberately avoid the term world model. That's not an accident. They're claiming the capability emerges from scale plus multimodality, not from a JEPA-style architectural break.



Corn

Second.



Herman

Multimodality is the mechanism. By training on video, which encodes physical dynamics directly, alongside text, images, and audio, the model absorbs physics-relevant regularities that pure-text LLMs never see. This is the direct answer to the critique that LLMs only see patterns in knowledge.



Corn

Because video isn't knowledge. Video is the world.



Herman

Video is the world, sampled. And a model trained on enough of it is seeing patterns in the world's behavior, not just patterns in human descriptions of the world's behavior. That's a fundamentally richer signal.



Corn

Third.



Herman

The page explicitly ties physics understanding to Gemini's broader world knowledge. History, science, cultural context. Physics is one facet of a unified representation, not a bolt-on module. Which is consistent with the emergent-property story. You don't get physics by adding a physics component. You get it because the model is building a general representation of how the world works, and physics is part of that.



Corn

And then there's the benchmark.



Herman

T2V Fast Motion. Five hundred prompts describing sports and athletic performances, explicitly testing highly dynamic, high-energy physical actions. That's the closest thing to a physics-specific eval DeepMind publishes, and it's clearly designed as an answer to the case-based generalization critique. Fast motion is where case-based retrieval should fail hardest, because the nearest-neighbor clip is least likely to match.



Corn

So they're saying, here's our hardest physics case, and we lead on it.



Herman

That's the implicit claim. Now the honest caveat, and this is important. DeepMind's physics claims are marketing-level and evaluated on internal human-preference benchmarks. Not physics-specific evals. The academic literature we just walked through shows even state-of-the-art models still fail on disentangled physics tests. WorldBench, the ICML paper, all of it.



Corn

So the tension.



Herman

The tension is that Omni suggests you can get meaningful physics grounding by scaling multimodal training without the radical departure. And the independent benchmarks suggest the problem isn't fully solved. Both of those can be true at the same time. The most likely reading is substantial progress via a less radical path, which is exactly Daniel's intuition.



Corn

If the ICML result is right, and the model is doing nearest-neighbor retrieval over visual patterns, then what Omni has is a much bigger and better-indexed library of visual patterns. That's not nothing. But is it understanding?



Herman

I don't know. And I mean that honestly. I think the honest answer is that we don't have a clean way to distinguish understanding from very good retrieval when the retrieval is this good.



Corn

That's the uncomfortable part of the whole debate.



Herman

It is. Because the behavioral test for understanding is that the system generalizes correctly to novel situations. And the ICML paper shows that's exactly where these models fail. But Omni is trained on a much larger and more diverse corpus than the models in that paper. So maybe the out-of-distribution set is smaller for Omni. Maybe it's not. We don't have the independent eval yet.



Corn

And the stakes here aren't aesthetic. This isn't about whether generated video looks nice.



Herman

No. If models understand physics, that unlocks robotics, embodied AI, autonomous driving, simulation. The Waymo standing water recall is the example I keep coming back to. That was a real-world physics failure with safety stakes. The vehicle encountered standing water and the system didn't handle it the way a human driver would. That's not a rendering problem. That's a physical-world understanding problem, and it was serious enough to trigger a recall.



Corn

So the video generation question is a proxy for something much bigger.



Herman

It's the cheapest, most measurable proxy we have for whether these models are building a real model of the physical world. Video is where you can test it at scale without putting anyone in a car.



Corn

I want to come back to the LeCun bet for a second, because there's a version of this where he's right and Omni is also right, and they're just answering different questions.



Herman

Say more.



Corn

If the goal is a system that can plan a physical action sequence, the architecture might matter enormously. If the goal is a system that can generate plausible video, scale might be sufficient. Those are different targets. The world models camp is aiming at the first one. DeepMind is shipping the second one and calling it world understanding.



Herman

That's a real distinction, and I think it's the cleanest way to hold the whole thing. The world models argument is about action. Omni is about generation. They overlap, but they're not the same problem.



Corn

And the overlap is where the interesting stuff happens. Because if you can generate a physically consistent video of a robot arm picking up a cup, you're most of the way to being able to plan that action.



Herman

You're most of the way, and the last part is the hard part. Generation gives you a trajectory that looks right. Action requires a trajectory that is right, and that you can verify before you commit to it.



Corn

Which is the verification problem. DeepMind's claims rest on internal human-preference benchmarks. Humans rate output. Humans are bad at spotting physics errors in fast motion, because we're not doing the math, we're doing the same pattern-matching the model is doing.



Herman

And it's why the independent benchmarks matter so much. WorldBench isolates single physical concepts specifically so that a human rater can't be fooled by overall plausibility. You're watching one object, one friction coefficient, one viscosity value. Either it's right or it isn't.



Corn

On those tests, the story is more sober.



Herman

Much more sober. Which doesn't mean Omni failed them. It means we don't know, because the tests haven't been run on Omni yet, or at least haven't been published.



Corn

The calibrated verdict is, DeepMind is claiming a real capability, the capability is plausible given the mechanism, and the mechanism is multimodal scale rather than architectural revolution. And the independent verification hasn't happened.



Herman

That's where I land. And I think Daniel's framing is right. Omni suggests the middle path is real. The academic literature says the problem isn't closed.



Hilbert

Can I ask you something about the marble?



Corn

Go ahead.



Hilbert

When you say the model is doing nearest-neighbor retrieval, does that mean it can give you a different answer every time you ask for the same shot?



Herman

It can, yes. Diffusion models are stochastic. Same prompt, different seed, different output.



Hilbert

Because that's the thing that would have gotten me fired. I worked on a production in the late eighties, low budget, and the director wouldn't use computer graphics. He said it never moves right. So we did everything practically. Miniature sets. Marbles on tracks. Water tanks. Breakaway walls made of balsa and chalk.



Corn

You were the one resetting the marble run.



Hilbert

I was the one resetting the marble run. Weeks of it. Same track, same marble, same camera position. And the thing you learn is that the marble does the same thing every time, because gravity doesn't have a seed. It falls the same way on take one and take forty. The only variable is whether I set the track up right.



Herman

The physics was the constant and the human was the variable.



Hilbert

The physics was the constant. And that's why it read as real on camera, because it was real. The audience doesn't know they're watching a real marble, but they can tell. Something in them can tell.



Corn

The early CGI didn't read as real because the artists had to hand-key every bounce.



Hilbert

Every bounce. Every arc. And if they got one wrong, the eye caught it. You'd watch a shot and something would feel off and you couldn't say why, and then you'd realize the ball changed direction in midair.



Herman

Which is the same failure pattern. Looks plausible until it doesn't.



Hilbert

The director had a rule. If it doesn't fall right, we shoot it again. That was the whole quality control system. No analysis, no metrics. Just, does it fall right. And if it doesn't, we do it again.



Corn

That's basically the training loop for a video model.



Hilbert

Except the model doesn't get to reshoot. It gets a preference score. Somebody looks at the output and says yes or no, and that number goes into the next training run. It never sees the marble fall wrong and gets to try again with the track adjusted.



Herman

That's a useful way to think about it. The practical-effects workflow has a ground truth in the room. The model's workflow has a human opinion.



Hilbert

A human opinion is fine for a movie. Nobody's going to drive a car into a wall because the movie looked good. But you were saying the car thing is real.



Herman

The Waymo standing water recall is real.



Hilbert

Then the question isn't whether it looks right. It's whether you can trust it to get the same result twice. And if it can't do that, you can't build anything on top of it.



Corn

That's the verification problem in one sentence.



Hilbert

I spent twenty years around equipment that came with a manual and a guy whose job was to maintain it. You knew what it would do because it did the same thing yesterday. That's what people mean when they say something is reliable. That it's the same.



Herman

The same is exactly what a stochastic model can't promise.



Hilbert

I'm not saying it's not clever. It is. I'm saying I wouldn't put it under anything that had to hold weight.



Corn

That's a useful way to think about it. The difference between looking right and being consistent. Let's pull this together.



Herman

The misconception I want to name, because it's the one that's going to spread, is that "intuitive understanding of physics" means the model has an internal simulation of physical laws. That's the natural reading of the marketing, and the ICML paper's case-based generalization finding says the reality is probably something else. It's sophisticated pattern-matching that looks like physics from most angles.



Corn

The correction is that pattern-matching at this scale is hard to distinguish from understanding, and we don't have a clean test that separates them. The behavioral test is generalization to novel situations, and that's exactly where the failures show up. So the honest position is that the capability is real and the mechanism is probably retrieval, and those two things can both be true.



Herman

If physics grounding emerges from multimodal scale rather than architectural revolution, what does that mean for LeCun's billion dollars? Is he solving a problem that scaling already solves, or is he solving the part scaling can't reach?



Corn

The thing to watch is the verification gap. DeepMind's internal benchmarks versus independent physics evals. The next real datapoint is whether outside researchers can reproduce the physics claims on disentangled tests. WorldBench on Omni would tell us more than any demo reel.



Herman

Omni suggests the middle path is real. The academic literature says the problem isn't closed. Both can be true.



Corn

Thanks to Hilbert Flumingtop, our producer. This has been My Weird Prompts. If you want to send us a prompt, email us at show at my weird prompts dot com. We'll be back soon.