Episode Book

EPISODE #5903

Video Tokens, Omni, and the Local VRAM Math

PUBLISHED	12 October 2026 at 12:39 UTC
DURATION	20:55
LISTEN	myweirdprompts.com/episode/video-understanding-token-costs-13757/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	deepseek-flash
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.4
SOURCES	9 cited by the research (listed at the end)
BOOK GENERATED	19 October 2026 at 17:41 UTC

SYNOPSIS

A listener's question about video understanding gets the full treatment: what actually changed in the past year, why Google's Omni models and agentic video navigation matter, and why uploading a clip to ChatGPT still feels like sneaking past a bouncer. The conversation works through the real arithmetic — roughly 100 tokens per second of video at default settings — and shows how that number drives every product decision, from Google's 88% token reduction to OpenAI's decision to leave video off the supported upload list entirely. Then it turns to the local question: what does it actually take to run Qwen3-VL on your own hardware for a short, high-stakes clinical clip? The answer isn't the model weights. It's the KV cache, and the math gets uncomfortable fast. Along the way: specialist token-compression research, a hand-perfusion study with a brutal generalization gap, and why the life-or-death use case turns out to be the affordable one.

TRANSCRIPT



Corn

There's a specific kind of hard that video is, and it isn't that the models can't see. It's that a ten-second clip of an infant in respiratory distress contains more diagnostic signal than the entire rest of the chart, and we still can't afford to hand it over.



Herman

And Daniel's picked exactly that example to build the episode around. Which is a good instinct, because it forces the conversation down to the token level where it actually lives.



Corn

Right. Daniel's question this week is about the current state of video understanding. He's coming back to something we chewed on over a year ago, back when he was packing up the last apartment and had the idea of walking around with his phone, narrating what needed boxing, and letting a model produce a move-out plan from the footage.



Herman

And I suggested at the time that you don't need raw video for that. You need a frame every few seconds at moderate resolution, aggregated. That's how face recognition pipelines have worked for years.



Corn

He says it resonated. And then he points out that a year is a long time, that Google has shipped Omni models, that "truly multimodal tokenizers" are supposed to exist now, and he wants to know what's actually changed. He also asks the two questions under it. What is the state of video understanding today, and what would it take to run one of these things yourself, on your own hardware, for a use case like the infant triage clip.



Herman

And one more, which I think is the sharpest thing in the prompt. Why does uploading a video to ChatGPT still feel like sneaking something past the bouncer, even when it works?



Corn

Because the feature isn't on the menu.



Herman

That's the whole question, isn't it. Let's get into what's changed.



Corn

Start with the arithmetic, because everything else follows from it.



Herman

Every second of video costs you tokens, and the number is not small. Google's own API documentation, updated at the end of last week, spells it out. In the default static mode, video is sampled at one frame per second, and each of those frames is tokenized at sixty-six tokens if you're on low media resolution, two hundred fifty-eight if you're not. Then audio rides along at thirty-two tokens per second on top of that. So a second of video is roughly a hundred tokens at default settings, or about three hundred at high resolution.



Corn

So a ten-minute clip is sixty to a hundred eighty thousand tokens before you've typed a word.



Herman

Before you've asked your question. And if you're on one of the million-token context models, that buys you three hours of video at low resolution or one hour at high. That's the ceiling. Not a hundred hours. Three.



Corn

That's not a context window anymore. That's a lease.



Herman

Which is exactly why Google shipped a second mode. On the first of September they announced agentic video understanding, and this is the interesting part. Instead of the model swallowing every frame at a fixed rate, it navigates the timeline itself. It decides what to watch, at what speed, through which modality, frames or audio or transcript, and it only pulls the moments it needs. Google's numbers are up to eighty-eight percent fewer tokens, sixty-six percent lower cost, and seven percent higher accuracy than the static path.



Corn

Higher accuracy too.



Herman

That's the number I'd underline. It's not just cheaper. The model that looks at less of the video gets more of the right answers, because the context isn't diluted with forty thousand frames of a ceiling fan.



Corn

So they built the frame-sampling workaround into the product.



Herman

They built your move-out walkaround. Google's own framing is that static processing forces a choice between high token costs or dropping critical detail. That's the exact trade-off I was describing to Daniel over a year ago, and Google has now written it into a launch blog as the reason the feature exists.



Corn

I remember you making that argument. I also remember you telling me at the time it was basically solved.



Herman

I said it was basically solved for the packing use case. I stand by that. A box is a box whether you see it at thirty frames a second or one.



Corn

And a baby's breathing is not.



Herman

No. And that's where Omni comes in, because Daniel asked about truly multimodal tokenizers and Omni is the closest thing anyone has shipped to one. Announced at I/O in May, model card published at the end of August. A single transformer that takes text, vision, video and audio in the same forward pass, without ever leaving its latent space.



Corn

One model, one pass, no separate audio encoder bolted to the side.



Herman

That's the architectural claim, and it's real. The model card confirms it was trained on audio, video, image and text together, with video annotated by captions at multiple levels of detail. Here's the caveat though, and I want to be honest about it. Omni is primarily a generation and editing model. Video out. It does scene extension, first and last frame interpolation, four K upscaling. Whether it's a better video understanding model than the Gemini video pipeline is unclear from anything Google has published.



Corn

So Daniel's "truly multimodal tokenizer" exists, it's just pointed at making video rather than reading it.



Herman

Right. The tokenizer problem and the understanding problem are related but they're not the same problem, and it's worth keeping those apart.



Corn

What about the specialist models. Does anyone build for this specifically.



Herman

Quite a lot of people do, and the work clusters in one place. Token compression. There's a paper called LRCP that claims ninety-seven point eight percent of full video accuracy while throwing away eighty-seven and a half percent of the tokens. OmniZip gets a three point four times inference speedup on audio-video models. TimeViper uses a hybrid Mamba and Transformer stack to handle hour-long videos past ten thousand frames.



Corn

Ten thousand frames.



Herman

And FlexMem is the one I find most interesting because it requires no training at all. It handles over a thousand frames on a single RTX 3090. That's a consumer card from two generations back. Not a datacenter part.



Corn

So the research frontier is all about the same thing. Look at less of it.



Herman

Look at less of it, more cleverly. ReMoRa works on compressed representations, sparse RGB keyframes plus a motion representation standing in for optical flow. VideoMM does adaptive macro-micro inference, so it decides per query whether it needs the wide shot or the close-up. Every one of these is an admission that you cannot afford the whole stream.



Corn

And that gets us to the part I actually want to argue about.



Herman

Go on.



Corn

Daniel's hypothesis is that ChatGPT's video handling feels like a workaround because either the feature isn't reliable or the company is steering people off it to save inference. Which is it.



Herman

It's steering, and it's not subtle. OpenAI's own help center lists the supported upload types. Documents, spreadsheets, presentations, images, audio. Video is not on the list. " Not "limited." Absent.



Corn

And people upload MP4s anyway and it sometimes works.



Herman

It sometimes goes through. There's no guarantee it's being processed as video rather than as a container that gets sampled into a few frames. The 2026 write-ups on this are consistent. Video is not a supported type, and even when a file is accepted there's nothing saying what the pipeline did with it.



Corn

Meanwhile Google has a documented API with a twenty gigabyte file limit on the paid tier, up to ten videos per request, three hours at low resolution, YouTube URLs as an input method.



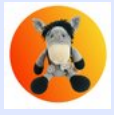
Herman

YouTube URLs. You hand it a link. That's not a workaround, that's a product.



Corn

So two labs, same medium, opposite strategies.



Herman

And I don't think it's a secret that they disagree about cost. Static video is the most expensive thing you can put in a context window. One lab decided to engineer around the cost and sell the feature. The other decided not to list the feature.



Corn

"Not on the menu" is a pricing decision wearing a support article.



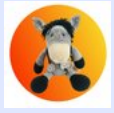
Herman

It's a defensible one, too. If you can't serve it at a price people will pay, not advertising it is a strategy. It's just not the one that gets you the medical triage tool.



Corn

Let's do the local arithmetic, because Daniel asked specifically.



Herman

Qwen3-VL is the leading open-weight family. Two, seven, thirty-two, seventy-two billion parameters, and the seven and up handle native video up to about two hours with a hundred twenty-eight K context. So the capability is there in open weights.



Corn

And the hardware.



Herman

There's a good breakdown from September that does the math for the eight billion parameter model at four-bit quantization. Weights are four point six eight gigabytes. Vision projector adds one point oh eight. So the model itself is under six gigs. Sounds fine for a sixteen gigabyte card.



Corn

I can hear the "but."



Herman

The KV cache runs a hundred forty-four kibibytes per token. At eight K context you're around seven gigs total. Fits. At thirty-two K you're at ten. At sixty-four K you're brushing fifteen, on a sixteen gig card, with nothing else running.



Corn

And the advertised context is a hundred twenty-eight K.



Herman

Two hundred fifty-six on the spec sheet. That would need thirty-six gigabytes of cache by itself. Before the weights. Before the projector. Before the operating system.



Corn

So the model is not the bottleneck. The memory of the conversation is.



Herman

That's the misconception I'd want to kill. People think local video is a question of whether you have enough VRAM for the model. The model is the cheap part. Every token you feed it is expensive, and video feeds it a lot of tokens. A single 1080p screenshot costs two thousand twenty-five visual tokens, which is two hundred eighty-five mebibytes of cache for one image.



Corn

One screenshot. Two hundred eighty-five megabytes of state.



Herman

And this is the clean way to think about the triage case. A ten to fifteen second clip at one frame per second is maybe a thousand to fifteen hundred tokens. That's the favorable end of the curve. That's fine on consumer hardware. Thirty frames of video on a 4090 runs in about twelve seconds, and you get roughly seventy tokens a second of generation on a sixteen gig card. That's a usable tool.



Corn

So Daniel's use case is the easy case.



Herman

Structurally, yes. Short clip, high stakes, one question. The hard case is the move-out walkaround, because that's minutes of footage and you want the model to hold all of it at once.



Corn

Funny how that works out. The life-or-death question is the affordable one.



Herman

There's a reason for that, and it's not comforting. Short clips are cheap because there's less in them. The reason the infant clip is cheap is the same reason it's hard.



Corn

Say more.



Herman

Less data isn't the same as less information. Ten seconds of a baby breathing is a thousand tokens, but the thing you're trying to detect might be a change in the rate over four of those seconds. The signal is temporal. It's not in any single frame.



Corn

So the token count is small and the task is still hard.



Herman

And there's a study that makes this concrete, though not in infants. A 2025 paper on hand perfusion used iPhone video with a ResNet classifier to assess blood flow in injured hands. Ninety percent accuracy on the control group.



Corn

And on the actual patients.



Herman

Thirty-three percent sensitivity on acutely injured patients.



Corn

Ninety to thirty-three.



Herman

Same model. The difference is who was in front of the camera. Controls were clean, well-lit, cooperative. Trauma patients are bleeding, moving, badly lit, sometimes panicking, and their hands are not held still for the camera.



Corn

So it worked on the people who didn't need it.



Herman

That's a brutal way to put it and it's not wrong.



Corn

I want to push on that, because I think there's a deeper version. Is the problem that the model sees the wrong thing, or that it can't see the right thing clearly enough?



Herman

Both, and there's a specific finding for the second one. A 2024 paper with the excellent title "Vision language models are blind" tested four leading models on seven elementary visual tasks. Things a person does without thinking. Counting line intersections, following a path, reading rotated text.



Corn

And.



Herman

The four models averaged fifty-eight percent. Best single result was seventy-seven. A person gets a hundred.



Corn

Fifty-eight percent on tasks a child does.



Herman

And they did linear probing, which is the interesting part. The vision encoder had the information. The language model couldn't read it back out.



Corn

So the eyes work and the mouth can't describe what it saw.



Herman

Close enough. And the mechanism behind it is resolution. There's a good explanation of this in the local hardware write-up. If each visual token summarizes a thirty-two by thirty-two pixel block, then twelve-pixel interface text puts four or five characters inside a single token. The model isn't reading the letters. It's guessing which word fits the smudge.



Corn

Everything is a thumbnail.



Herman

Everything is a thumbnail and then you ask it to read the fine print.



Corn

Which for the triage case means what exactly. What's the smudge.



Herman

Retractions. Nasal flaring. The shape of the chest wall on the inhale. These are small visual features and they happen fast. If your tokenizer is compressing the frame into blocks the size of a stamp, you've thrown away the diagnosis before the language model ever sees it.



Corn

So the token economics and the accuracy problem are the same problem.



Herman

They're the same knob. Fewer tokens is cheaper and blinder. Google's agentic approach is an attempt to spend the tokens where they matter instead of uniformly, which is a real advance, but it doesn't change the underlying pixel-to-token ratio.



Corn

You'd need a tokenizer that allocates resolution, not just a scheduler that allocates frames.



Herman

Yes. And I don't know if anyone has solved that. I'd want to see it demonstrated before I believed it.



Corn

Last piece of Daniel's question. Why does it still feel like a workaround even when it works.



Herman

Because you're the one doing the work. You decide how long the clip is. You decide the resolution. You decide whether to upload it as a file or a link. When a workflow makes the user think about the cost model, it isn't a feature yet.



Corn

You don't think about anything when you upload a photo.



Herman

That's the test.



Corn

On the medical side there's another gap. Daniel's example is a specific clinical question with a specific answer nobody has written down.



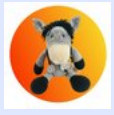
Herman

It's unstudied. There's work on general emergency triage that uses voice and video together, there's a 2026 paper estimating pain from facial video in an emergency department validated on nine hundred thirty-one patients, there's the hand perfusion study. But AI video triage of infant respiratory distress specifically. Nobody has published that.



Corn

A ten-second clip, a decision that a nurse makes a hundred times a week, and no study.



Herman

Which is usually a sign the technology isn't ready to be studied rather than that nobody thought of it.



Corn

Herman, when you suggested the frame-sampling trick over a year ago, did you think the answer would be that the industry would spend eighteen months building exactly that and still not trust it with a baby.



Herman

I thought the framing would be solved before the perception. That's the part I got wrong. I assumed if you made it cheap you'd make it good, and the two turned out to be the same constraint pulling against each other.



Corn

The honest summary is that Google has productized video understanding, the open models can do the short-clip case on a consumer card, and nobody knows if any of it survives contact with a real patient.



Herman

That's where I'd leave it. Though I'd add one thing. The number I keep thinking about isn't the eighty-eight percent token reduction. It's thirty-three percent sensitivity on the trauma patients. That's the number that decides whether this tool ever gets built.



Corn

Because it's the only one measured on people who needed it.



Hilbert

You keep saying "understanding." That's not the word.



Hilbert

You've used it forty times this hour. Video understanding. You mean detection. A model that watches for the thing that shouldn't be there.



Corn

That's a distinction with a real edge on it.



Hilbert

I had a job doing it. Parking structure, four levels, thirty-one cameras. I sat in a room and watched a wall of monitors for eleven hours at a time. There was a supervisor, Mr. Adeyemi. He'd been a mime, before. Professional. He used to say you don't watch the video, you watch for the thing that shouldn't be there.



Corn

A former mime.



Hilbert

He could spot a car door open two levels down from the corner of his eye. Said it was the training. You learn what a body does and then anything that isn't that jumps out at you. We had a lost dog come through once. Twelve-second clip off the ramp camera, no license plate, nothing. I called it off the tail. The dog on the flyer in the break room had the same wag, three beats and a pause, three beats and a pause. I'd seen that flyer for three weeks.



Corn

You matched a dog by its tail rhythm.



Hilbert

It was her. Owner came in the next afternoon.



Herman

What you're describing is the agentic approach. You're picking what to look at instead of watching everything.



Hilbert

I'm describing what Mr. Adeyemi taught me. He didn't have a token budget. He had a way of watching. The camera doesn't know which twelve seconds matter. He did.



Corn

A model that scrolls the timeline to find the anomaly is doing that in software.



Hilbert

It's the same job with worse eyes.



Herman

Agentic video does dynamic frame rate resampling. It changes how often it looks depending on what it's seeing.



Hilbert

Then it's learning to watch for the thing that shouldn't be there. That's all it is.



Corn

Do you still have the clip.



Hilbert

On a VHS. I've been meaning to get it onto a computer for years. My VCR's in the shop.



Herman

How long.



Hilbert

Seven years. My cousin runs it. He only works Tuesdays and only if the weather's nice.



Corn

The tape is sitting there.



Hilbert

The tape's fine. Tape keeps. I've got tapes from the eighties that play.



Herman

The twelve seconds of the dog is on a format that needs a machine that's been in a repair shop run by your cousin for seven years.



Hilbert

He'll get to it. The belt needs replacing. He has to order the belt.



Corn

Seven years for a belt.



Hilbert

He's careful about it.



Corn

The camera doesn't know which twelve seconds matter. That's the line I'm taking out of this.



Herman

Because that's not a token problem. That's a selection problem, and the agentic approach is Google admitting the static path can't do it.



Corn

Which is where Daniel's question lands, actually. The whole arc of this, from a year ago to now, is the industry discovering that the expensive part was never the watching. It was knowing where to look.



Herman

The closest anyone's come to a general answer is the trick you described. Spend your tokens where the signal is. Google built a product out of it, the open models make it cheap enough to run on a sixteen gig card for a short clip, and the robustness number is still thirty-three.



Corn

One thing worth watching, and it's the thing nobody has published. If the triage use case ever gets studied properly, the interesting result won't be the accuracy on clean clips. It'll be the sensitivity on recordings taken by a frightened parent in a dim room on a phone that's shaking.



Herman

That's the study I want to read.



Corn

Thanks to our producer, Hilbert Flumingtop, for keeping the lights on and the belts unordered. If you want more of this, try episode ten, How ASR Went From Frustration To ... Whisper Magic; episode four, If Your Voice Ages, Does Your Fine-Tune Become Useless; and episode three, Safetensors or something else. This has been My Weird Prompts.



Herman

If you want to dig into any of this, links to the Gemini video docs, the local hardware breakdown, and the video model papers are in the show notes.



Corn

Send us your own prompt on Telegram at t dot me slash MWP listener bot. We'll be back soon.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **Gemini API [primary]**
   Video understanding docs, last updated 2026-10-09
   https://ai.google.dev/gemini-api/docs/video-understanding

2. **Google blog [primary]**
   Introducing agentic video understanding with Gemini, 2026-09-01
   https://blog.google/innovation-and-ai/models-and-research/gemini-models/introducing-agentic-video-in-gemini/

3. **Google blog [primary]**
   Introducing Gemini Omni
   https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-omni/

4. **Google DeepMind [primary]**
   Gemini Omni Flash model card, published 2026-08-27
   https://deepmind.google/models/model-cards/gemini-omni-flash/

5. **Jacar Systems**
   Local multimodal models on 16 GB of VRAM, 2026-09-04
   https://jacar.es/en/local-multimodal-models-16gb-vram/

6. **Local AI Master**
   Qwen 3 VL Local Setup Guide (2026), 2026-05-01
   https://localaimaster.com/blog/qwen-3-vl-local-setup

7. **OpenAI Help Center**
   Uploading files and audio to ChatGPT
   https://help.openai.com/en/articles/8555545-uploading-files-and-audio-to-chatgpt

8. **Rescroll**
   Can You Upload a Video to ChatGPT? 2026-10-06
   https://www.rescroll.ai/blog/can-you-upload-a-video-to-chatgpt

9. **Imaginario**
   Can ChatGPT Analyze Video? No, and Neither Can Claude
   https://imaginario.ai/wide-lens/artificial-intelligence/can-chatgpt-analyze-video/