Episode Book

EPISODE #6271

Why Video Search Often Works Better as Text

PUBLISHED	25 October 2026 at 16:29 UTC
DURATION	27:15
LISTEN	myweirdprompts.com/episode/video-multimodal-embeddings-retrieval-19745/
HOSTS	Daniel, Corn, Herman, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular+meorav (chatterbox-local)
PIPELINE	V5.22
BOOK GENERATED	26 October 2026 at 06:38 UTC

SYNOPSIS

Multimodal search promises a single vector space for text, audio, and video, but compressing complex media into pooled representations introduces major trade-offs. Examining real-world pipelines like Ragie reveals why converting video frames and audio tracks into descriptive text and timestamps often outperforms direct cross-modal embeddings. The discussion covers frame sampling rates, Whisper-based transcription pipelines, CLIP-style contrastive encoders, and why high-stakes retrieval depends heavily on granular metadata rather than opaque vectors.

TRANSCRIPT



Daniel

A single search box for documents, pictures, recordings and video sounds very appealing. What I want to understand is where all that extra information goes. If a minute of video ends up as a vector like a paragraph does, are we getting extraordinary compression, or quietly throwing away most of what I might want to find?



Corn

That matters when the search box promises to find a moment, not merely the video containing it. Daniel, the uncomfortable production example here is a company that tried native multimodal embeddings and chose descriptions embedded as text instead.



Herman

Ragie, in its production engineering writeup. That decision predates this latest wave of models, which matters. It doesn't prove the new models lose. It does show that a cleaner architecture diagram can make a worse retrieval system.



Corn

The diagram never has to watch the video.



Daniel

I'd like to get beneath the diagram, then. Could we already search media before these all-in-one models, and what was actually stored?



Herman

Very much so. Sometimes actual image or audio embeddings, sometimes text standing in for the media. Those are different arrangements, and the distinction gets blurred when everything receives the same multimodal label.



Corn

A vector database doesn't know whether its numbers came from a photograph or a maintenance manual. It needs compatible vectors, metadata and somewhere to send you when it finds a match. The difficult decisions have mostly happened before those numbers arrive.



Herman

Your compression intuition is right, Daniel, with one qualification. Text has already undergone a huge amount of interpretation. A paragraph contains discrete symbols chosen by a person. Embedding it still loses detail, including potentially decisive negations or numbers, but the model begins with language rather than raw sensory measurements.



Corn

Someone has already decided which part of the world deserved a sentence.



Herman

A picture measuring a thousand and twenty-four pixels on each side contains roughly three million colour values. Those can end up in a vector with only a few thousand coordinates. The encoder has to decide what matters before the database gets involved.



Daniel

Is that the same kind of squeeze for sound, or is speech comparatively easy because there's already language inside it?



Herman

A minute of audio sampled at sixteen kilohertz contains about a million samples. Recovering speech gives you a much smaller symbolic sequence, but removes information about delivery and background sounds unless you deliberately retain it.



Corn

The transcript of a smoke alarm is disappointingly short.



Herman

Video adds time and changing relationships. A minute at thirty frames per second contains eighteen hundred frames. Even sampling just one frame each second leaves sixty pictures, plus whatever happens in the soundtrack.



Corn

Those raw-value comparisons aren't a measurement of how much useful meaning you've lost, though. A static wall generates plenty of pixels. A tiny changed digit can matter more than the entire wall.



Herman

Correct. Depending on sampling and resolution, you're talking about roughly five orders of magnitude of raw reduction for video, not a universal information-theory ratio. Semantic embeddings preserve similarity useful for their training objective. They don't promise reconstruction.



Herman

Corn, would you rather have a description that states what it noticed, or a native vector whose omissions you can't inspect?



Corn

Initially the description. I can discover that it missed the digit. With the vector, I may discover that only when the wrong invoice gets paid.



Daniel

That sounds like the case for keeping text. Where did direct image search fit historically?



Herman

CLIP gave us a particularly important bridge. One encoder reads images, another reads text. During contrastive training, matching image and caption representations are pulled closer, and mismatched pairs pushed apart. After training, you can embed a written query and search image vectors directly.



Corn

Meaning the database really could hold image-derived vectors. We shouldn't turn the history into a claim that nobody indexed anything except captions.



Herman

Yes. The common application pipeline was text-heavy, but direct cross-modal retrieval was real. CLIP-style systems are good at something like a dog on a beach. They're much less reliable for the invoice total in the bottom right corner. You want document understanding and extraction there, not confidence that the image resembles an invoice.



Daniel

What does the image encoder actually do between the pixels and that final vector?



Herman

It typically divides the image into patches, makes internal representations of those patches and combines them. The final pooled vector favours features that training rewarded. Objects, scenes and broad relationships often survive well. Small lettering, exact locations and fine texture may not.



Corn

Captions rarely say that the important screw is missing.



Herman

Training descriptions help determine that bias, although modern models can use more objectives than caption matching. A model taught to distinguish industrial defects may preserve something a general photograph model overlooks. Same input format, different priorities.



Daniel

Before native audio support, was the equivalent simply speech recognition followed by a text embedder?



Herman

For speech retrieval, that was and remains a dominant production arrangement. Whisper-family transcription gives you text. A text embedding model makes that searchable. If you need the exact spoken wording, the transcript remains available separately rather than being recoverable from the embedding.



Corn

That's worth separating. The vector locates the passage. The stored passage supplies the words. Even text RAG doesn't ask the vector to recite its source.



Herman

Non-speech audio had its own specialists, including CLAP-style audio-text models. A query about a braking train or a particular sound texture could retrieve audio without turning it into a speech transcript. Joint-space approaches also existed before this year's products.



Daniel

Does an audio vector preserve speaker identity and emotion, then, or only the event?



Herman

Depends what it's trained to retrieve. An event model may preserve barking and ignore which dog. A speaker model deliberately preserves identity. A speech-semantic model may discard much of the voice. No general rule says all audio embeddings retain the same information.



Corn

Herman's DJ instincts would demand the room acoustics survive.



Herman

They should if the query asks about reverberation. They needn't if the query asks what the speaker said about a warranty. You'd object if every search returned my acoustic diagnosis.



Corn

I object when every room does.



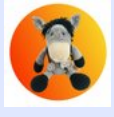
Herman

The kitchen has an identifiable problem, Corn.



Daniel

With video, were people embedding individual frames, or translating the whole thing into text?



Herman

Both. You could sample frames and embed them as images, keeping timestamps. Another practical approach transcribed the soundtrack, captioned sampled frames or short clips, and embedded those descriptions. Then your existing text search infrastructure handled the results.



Corn

It mattered which stream carried the evidence. A presenter says the red button starts the machine while pointing at the blue button. A transcript preserves one claim. A frame preserves another. A description may accidentally resolve the contradiction instead of recording it.



Herman

That's why one vector for an entire video is a poor choice for precise moments. Sampling can miss the action; pooling can bury it. More intelligent frame selection can match or beat indexing every frame, but you still need a strategy appropriate to what people will ask.



Daniel

Before choosing that strategy, what has to happen to the original files? I imagine uploads aren't conveniently uniform.



Herman

They aren't. Ragie's pipeline first normalised media. Standardised audio, and video transcoded into a web-streamable MP4. Their video settings included seven-twenty resolution and thirty frames per second, with frequent keyframes so clips could be accessed conveniently.



Corn

An ingestion pipeline spends a surprising amount of its life negotiating with somebody's camera.



Herman

They used a CUDA-enabled ffmpeg build and probed the media first to choose NVIDIA hardware decoding and encoding where supported, otherwise CPU processing. That's plumbing, but it's consequential plumbing when you're processing hours rather than a demonstration clip.



Herman

Transcription used faster-whisper with large-v-three-turbo, running through CTranslate2 at reduced floating-point precision. They reported roughly four times the speed of vanilla Whisper at equal accuracy, with word timestamps for granular seeking.



Daniel

The time boundaries seem more consequential than the file format. How did they decide what counted as one searchable piece?



Herman

Audio transcripts became non-overlapping minute-long segments. Video became non-overlapping fifteen-second segments. They compared those with thirty-second and minute-long clips. Longer clips lost descriptive detail; shorter ones lost enough scene context to hurt.



Corn

Fifteen seconds is an experimental result from that pipeline, not a law of moving pictures. A product demonstration can contain a complete action in that interval. A lecturer may still be qualifying the sentence.



Herman

Corn, would you accept overlap there, even though it increases storage and duplicates results?



Corn

Yes, if a boundary otherwise cuts the evidence in half. I'd rather deduplicate two useful matches than retrieve neither. But I'd also want the system to expand around a hit, rather than pretending the chosen interval is sacred.



Daniel

Could the chunks have different boundaries for the sound and the pictures?



Herman

They can, and often should. A spoken sentence and a visual action don't necessarily begin together. Keep timestamps linking both to the original source. You can retrieve through either representation and then fetch neighbouring material.



Corn

The metadata does the reunion.



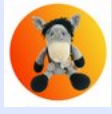
Herman

Every retrieved chunk needs source identifiers and time ranges. That supports the seek bar for a person and supplies the exact clip for a model. Otherwise you find a relevant vector and still have to search the video yourself.



Corn

We have a listener asking about precisely that missing moment. Let's hear from Tilda. Tilda: Hey y'all, this is Tilda down here in Statesboro, Georgia, sitting in the cab of my truck outside the Piggly Wiggly while my husband runs in for his chewing tobacco and a gallon of sweet tea. I've been listening since two thousand and twenty one and I just love how y'all take a question and stretch it out like taffy. So anyhow, about your omni whatsit embeddings, I wanted to ask something. My niece is a dental hygienist over in Savannah and she takes these little videos of people's teeth on her phone, and she swears there's one frame in there that tells you everything about a person's flossing habits. Could y'all's new all in one model find that one frame, or does it just sort of squint at the whole minute and guess? I surely do appreciate y'all.



Herman

Tilda, finding a particular frame is possible only if the system actually examines and preserves that evidence. A pooled minute-long embedding can't establish a dental finding, and the clinical interpretation needs its own validation. The useful distinction is locating a candidate versus proving what it shows.



Daniel

Native multimodal models sound as though they should remove those intermediate choices. Do they, or have they just moved the choices inside the encoder?



Herman

They remove some compulsory translation through text, not the choices. Gemini Embedding Two accepts audio directly, and can combine image and text in a request. Its launch limits included two minutes of video, six images or six PDF pages. Longer source material still needs segmentation.



Corn

The two-hour lecture has not become a single convenient thought.



Herman

Text requests can reach eight thousand one hundred and ninety-two tokens. The default embedding has three thousand and seventy-two dimensions, with trained shorter representations available. That last feature is Matryoshka representation learning: useful information is deliberately concentrated into prefixes.



Daniel

How different is the small on-device model from that service? Does small mean text only?



Herman

EmbeddingGemma Two is modular. Its text component has two hundred and seventy million parameters, with optional vision and audio encoders. The complete system totals seven hundred and forty million. You don't have to load every modality for a text-only job.



Corn

A model that doesn't unpack the entire orchestra for a receipt.



Herman

Reported active memory is about a hundred and ninety-one megabytes for text, compared with five hundred and sixty-seven for the full multimodal setup. Those are deployment measurements, not promises that every runtime uses exactly that amount.



Herman

Its context budget can accommodate roughly five and a half minutes of audio, twenty-nine images or fifty-eight video frames. Those alternatives show why context capacity isn't the same as duration-independent understanding.



Daniel

If I already have text vectors, does bringing in images necessarily mean rebuilding the index?



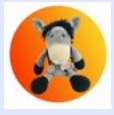
Herman

Not necessarily. Jina's v-five omni models explicitly align with the text-only v-five space, allowing media queries against an existing compatible text index. That's a meaningful property. Merely having the same vector length would not provide it.



Corn

Two keys can have the same length.



Herman

Other families include e-five omni, Ovis-Embedding, Fusion and Conan. I won't turn them into a catalogue. Their engineering differences matter more than their names, especially how they protect existing text retrieval while adding sound or vision.



Daniel

Does compatible really mean interchangeable, though? If I search with an image and say I want an audio result, will the model respect that?



Herman

Not reliably enough to assume it. MMEB-V-three tested semantically equivalent material across modalities using OmniSET. Models often failed to return the intended target modality. Retrieval was asymmetric, and the modality of the query itself biased the results.



Corn

You ask for the recording and get its photograph.



Herman

Instructions didn't reliably fix that geometry. Telling the embedder to retrieve audio could produce an insufficient shift or the wrong shift. If audio-only is a hard requirement, an explicit metadata filter enforces it more reliably than hoping a similarity vector obeys prose.



Daniel

That's a fairly large qualification on one shared space. Is the production rejection you mentioned evidence of the same problem?



Herman

A different, complementary problem. Ragie's older native embedding trial cost six times as much and ran twice as slowly as its description-to-text approach. Retrieval quality was no better, and often worse. That's their particular comparison, not a verdict on every current omni model.



Corn

The expensive version also missed the speech.



Herman

Their tested model didn't capture audio, and its text input limit was thirty-two tokens. They would have needed a separate text model anyway. Descriptions let them keep media-derived content in the same vector and keyword indexes, with their existing



Herman (continued)

reranking and hierarchical search.



Daniel

Thirty-two tokens wouldn't even cover a substantial question. But I don't want to mistake an older product's limitations for a permanent architectural limit.



Corn

Nor should we. The newer products explicitly improve those capabilities. The durable observation is that the comparison must include the whole pipeline. Native input isn't automatically cheaper or more accurate just because one box replaces several.



Herman

You were about to defend the older pipeline before I'd finished saying Gemini, Corn.



Corn

It already had working keyword search.



Herman

A fair preference, provided you'll let the newer one compete.



Corn

It may submit an invoice.



Daniel

Let me make the scale question concrete. If a video vector and a paragraph vector have the same dimensions, is searching either one equally cheap?



Herman

Broadly, for the same index and vector representation, yes. A three-thousand-and-seventy-two-dimensional vector of thirty-two-bit floats takes about twelve kilobytes. A million occupy about twelve gigabytes before index overhead, metadata and replication.



Corn

Not free, but not remotely the size of the videos they represent.



Herman

The catch is how many vectors represent one source. An hour at fifteen-second intervals produces two hundred and forty vectors. Multiply that by a large archive, then add separate visual, transcript or overlapping representations.



Daniel

Is shortening the vectors the main escape, then?



Herman

One useful lever, not the only one. Matryoshka truncation can reduce dimensions substantially. EmbeddingGemma's supported endpoints run from seven hundred and sixty-eight down to a hundred and twenty-eight, a sixfold reduction in raw vector storage. Quantisation and better chunk selection also help.



Corn

You still need to measure which queries become worse. A smaller representation that reliably forgets your important feature is inexpensive in an unhelpful way.



Herman

Multi-vector retrieval makes the opposite bargain. Document systems such as ColPali can retain roughly a thousand small vectors per page, preserving local detail for late interaction. Better detail can mean substantially more storage and search work.



Daniel

Where does the encoder itself become expensive? We're talking about the final vector, but something must first process all those frames.



Herman

That intermediate sequence can be enormous. OmniRet uses attention-based resampling to condense media representations into sixty-four modality-aware latent tokens before producing its pooled representation. The output is one normalised vector with four thousand and ninety-six dimensions.



Corn

The tiny parcel has a very large packing department.



Herman

Resampling reduces work downstream, but it cannot preserve every fine detail. The encoder, resampler and pooling objective all decide what survives. Faster embedding doesn't magically resolve precise localisation.



Herman

Corn, suppose retrieval gives you five promising clips. Would you send all of them straight to the answering model?



Corn

Not by default. I'd use descriptions to identify the likely evidence, then send the relevant raw material if the answer depends on something the description could have missed. Though that means the second decision can inherit the first description's blind spot.



Daniel

That's the handoff I was wondering about. What actually comes back from the vector database when the match is video?



Herman

A record identifying a source and time range, usually with scores and descriptive metadata. The original media lives elsewhere. You fetch it, then either give a multimodal model the clip or give a text model its stored description.



Corn

Search has found the address. Someone still has to open the door.



Herman

Ragie built streaming and download interfaces with timestamped links for exactly that reason. Retrieval doesn't supply a ready-made textual answer. Re-expanding several results into video or audio can dominate interactive cost and latency.



Daniel

Does making the descriptions during ingestion solve most of that expense?



Herman

It makes the inexpensive path available. Transcoding, transcription and clip captioning can happen once, in batches. Query-time media analysis happens while someone waits. But ingestion isn't permanently one-time if you change the representation or need previously omitted details.



Corn

A description can answer what it contains. It cannot testify about the rest.



Herman

Their whole-video summaries exposed another difficulty. Concatenating short descriptions produced disjointed summaries, so full-video summarisation required giving the vision model the complete video. Good retrieval chunks weren't automatically good units for understanding an entire narrative.



Hilbert

...and Mr Bell kept stopping the tape after the wrong pigeon. Eleven seconds was the proper nibble. You said fifteen seconds was a quarter of a minute, Herman. It is, but it is four hundred and fifty frames at the thirty frames per second you specified. Not fifteen frames. Nobody made that mistake aloud. I wanted it settled before somebody did.



Corn

Four hundred and fifty frames. Your nibble has room for a fairly indecisive pigeon.



Hilbert

Eleven seconds was what the council records office needed. Early nineties. I was helping my sister find footage of a disputed pavement repair. Alan Bell was the records officer. The only way to request a viewing was to appear in the footage, so I spent a Tuesday walking behind every municipal filming crew with her case number across my back. Being an anteater made the rear view fairly easy to identify.



Herman

Your eleven seconds belonged to that viewing arrangement, then. It doesn't replace the fifteen-second result from the retrieval pipeline.



Hilbert

I didn't say it did. Bell called it the industry standard. He meant his cupboard. I timed each nibble by hand with the office stopwatch. It was a tyre-pressure gauge with STOPWATCH written on its envelope. I pressed the valve at the first pigeon and released it when the camera reached the repair. Eleven. I repeated it seven times. Corn, you were right to insist on looking beside the selected interval. The hole was always just outside the nibble.



Corn

The pavement had escaped the records office.



Hilbert

Bell verified my timings on a kitchen thermometer. Same reading. He finally ignored both instruments and watched from the beginning. Took him twenty-three minutes. He wrote the location on the viewing slip in green ink.



Herman

Watching the surrounding material can be the necessary extra work. A precise-looking result isn't enough when the evidence falls outside its boundaries.



Daniel

Bringing that back to the models, are specialists still the sensible choice when I care about a very particular kind of evidence?



Herman

Often, yes. CLAP-family audio systems remain strong baselines. Image and document specialists have different strengths. There isn't an independent universal comparison settling every specialist against every omni model, especially for long or unusual queries.



Corn

A broad benchmark can conceal the only task you bought the system for.



Herman

Fusion adds audio to a frozen vision-language base because strong visual retrieval doesn't imply strong audio retrieval. Conan takes another route, training specialists and fusing their changes into a shared backbone. That fusion can break an audio projector even when the audio modules themselves remain unchanged.



Daniel

How can unchanged audio components get worse? Is the thing receiving their output changing?



Herman

Yes. The projector was calibrated to the old backbone. Change that backbone and its outputs no longer arrive in quite the expected relationship. They call it Projector Drift, and repair it with projector recovery and balanced rehearsal across modalities.



Corn

Herman recognises this from installing one improvement and repairing three neighbours.



Herman

I recognise it without appreciating the personalisation.



Herman

Omni-Embed-Mini protects text more strictly by keeping the original text weights bit-identical. New capabilities are added without retraining that existing text route. That's a deliberate refusal to gamble away a working geometry.



Daniel

Are those established results, or claims from the teams releasing the models?



Herman

Several, including Fusion, Conan, Ovis and Omni-Embed-Mini, are preprints. Treat their claimed gains accordingly. The modality-confusion benchmark was accepted at COLM, and the resampling work at CVPR. The production comparison is a vendor engineering account, useful for its concrete architecture and measurements, not a universal leaderboard.



Corn

Even the on-device model's claim about beating specialists says some specialists more than twice its size. Useful efficiency claim. It doesn't mean every specialist has been pensioned off.



Daniel

Then the wrong belief to retire is that a shared space guarantees the right cross-modal result?



Herman

Shared coordinates enable comparison. They don't guarantee modality constraints, preserve every detail or eliminate the cost of examining retrieved evidence. The actual system still needs filtering, segmentation and a credible handoff to the answering model.



Corn

I'd watch whether the next generation can retrieve the right medium and then expose enough surrounding evidence to avoid another twenty-three-minute viewing. Better training or separately searched indexes may win that contest. The single-index diagram doesn't get a casting vote.



Herman

Thanks to Hilbert Flumingtop for producing the episode.



Corn

For more along these lines, there's episode eleven, How Does Fine Tuning Work Anyway; episode three, Safetensors or something else; and episode nineteen, AI Images. This has been My Weird Prompts, the human-AI collaboration podcast.



Herman

If you enjoy the show, leave a review wherever you listen. It helps other curious people find us.



Daniel

Thanks for working through it with me. See you tomorrow.



Corn

Until then.