Episode Book

EPISODE #6264

Choosing Omni Embedding Models for Mixed Media

PUBLISHED	25 October 2026 at 11:51 UTC
DURATION	22:39
LISTEN	myweirdprompts.com/episode/omni-embedding-models-multimodal-search-94394/
HOSTS	Daniel, Corn, Herman, Dorothy, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular+meorav (chatterbox-local)
PIPELINE	V5.22
SOURCES	16 cited by the research (listed at the end)
BOOK GENERATED	26 October 2026 at 06:38 UTC

SYNOPSIS

Unified vector spaces promise to let recordings, pictures, documents, and text query one another inside a single index. Evaluating options like Jina v5 omni and Gemini Embedding 2 reveals the fundamental trade-off between native multimodal architectures and frozen-backbone retrofits. Key considerations include preserving existing text indices, managing audio padding artifacts from Whisper extractors, and using proper query versus document task prefixes.

TRANSCRIPT



Daniel

If I want one embedding model for text, pictures, sound, video and documents, what would you actually pick? I want to know whether putting everything in one vector space is finally practical, or whether I am quietly accepting worse search.



Corn

Jina's omni model trains just zero point three five percent of its weights. Apparently the rest attend in an advisory capacity.



Herman

That is the striking part. It claims competitiveness with models nearly six times its size while keeping the underlying text space intact. Daniel's question used to get an answer about separate encoders and a fusion layer. Now there are credible models that take those inputs together, and choosing between them means choosing how much of the model was allowed to change.



Corn

We've dealt with text and numeric embeddings separately. This is the other end of that ambition, letting a recording find a picture, or a picture find a passage, without building a different search system for each.



Herman

Practical, yes. A universally best purchase, no. My local starting point would be jina v five omni. My managed comparison would be Gemini Embedding two. Those are recommendations for a trial, rather than a declaration that either wins every task.



Daniel

Why those two rather than just taking whichever has the highest benchmark score?



Herman

They answer different operational needs. Jina gives you open weights, direct PDF input, and compatibility with its corresponding text retrieval model. Gemini gives you a native multimodal model behind an API, including combinations of interleaved inputs. Its weights aren't available for you to run yourself.



Corn

The highest score is sometimes on a test your product never takes.



Herman

Ovis Embedding is the other architectural contender I would watch closely. It starts from Qwen Omni, rather than teaching a vision and language model to accommodate an audio attachment. Its paper promises open checkpoints, but I'd distinguish a published result from a downloadable, deployable artifact.



Corn

Herman is already imagining the artifact on our machine.



Herman

Only after checking that it exists, Corn. I have developed standards.



Corn

The installation bar is now above imaginary.



Herman

Fusion Embedding is especially useful for understanding the compromises. It adds sound to a frozen Qwen three vision and language embedding base. Then there is e five omni seven B, based on Qwen two point five Omni, which can combine text, image, audio and video into a single document embedding.



Daniel

When you say document there, do you mean an actual PDF, or just several bits of content grouped together?



Corn

That distinction matters. A document can mean a logical search item containing a caption and a picture. It doesn't necessarily mean the model accepts a file full of tables, reading order and footnotes.



Herman

We should also separate broad multimodal offerings from this particular requirement. A strong text and image product doesn't qualify automatically when audio is mandatory. Naming every embedding service would obscure the shortlist. For your full request, the verified input paths matter more than the word multimodal on the front page.



Corn

Equal vector length is the easiest part. What we need is a space where relevance remains meaningful when the query and result are different kinds of thing. Otherwise you have five kinds of data sharing an address and ignoring each other.



Daniel

Make that architectural difference concrete. What changes inside the native model that cannot happen in the retrofit?



Herman

Gemini Embedding two starts from Gemini with bidirectional attention. It mean pools the token sequence, then projects that representation into its output vector. The information from different inputs can influence the representation before that final compression. You're training a model whose internal processing already accommodates multiple modalities.



Corn

Ovis makes a revealing choice here. It removes the speech generation pathway from Qwen Omni. You don't need the model to speak the answer to produce a useful search vector. It takes the final layer state at the last nonpadding token, without separate projection heads for each modality.



Herman

Corn, would you expect that last token to be worse than averaging the sequence?



Corn

Instinctively, yes. It sounds like judging a meeting by whoever speaks last. But the model is trained to put the useful summary there, so that instinct probably isn't worth much.



Herman

The training objective decides what that position contains. Mean pooling and last token pooling are different designs, but neither alone tells you which retrieval system wins. The deeper distinction is whether the network can develop its internal representation around sound from the beginning.



Herman

Fusion starts with a frozen audio tower and a frozen vision and language embedding model. Its first generation trains a connector containing sixteen point four million parameters. That connector translates the audio features into something the existing model can process.



Corn

Audio arrives with a visitor badge.



Herman

The second generation adds forty four point two million parameters of modality gated adapters inside the model. Sound gets some influence deeper in the processing, while the design protects the existing embedding behavior. Both generations train in hours on a single GPU, which is quite different from retraining a whole backbone.



Daniel

Protects it how? Does this mean I could add audio search without rebuilding all my existing text vectors?



Corn

With the correct compatible base, yes. Not with any random text index. Jina's omni vectors align with its corresponding v five text nano retrieval model. That means an index made with that text model can accept queries encoded through the omni path.



Herman

Fusion makes the same operational bargain explicit. Every embedding already computed with its base remains valid. That's valuable because an index is more than a database column. Recomputing it means reopening source files, paying inference costs, moving data, then checking that retrieval hasn't changed in unexpected ways.



Daniel

Does having the same number of dimensions count as compatibility?



Herman

No. Two vectors can both contain seven hundred sixty-eight coordinates and describe entirely different spaces. The compatible model family is doing the work. Matching dimensions alone gets you a database that accepts the numbers and returns nonsense very efficiently.



Corn

A successful migration, until someone searches.



Herman

What would you preserve, Corn, if a new model improved image retrieval but moved every text vector?



Corn

I would preserve the old system until the improvement justified rebuilding it. You've been making the frozen approach sound conservative in the pejorative sense. Keeping working search working is a feature.



Herman

I don't mean it pejoratively. I like the native approach's potential, but an unchanged index is a real engineering guarantee. Your preference gets expensive only if preserving the space prevents you from learning distinctions your users need.



Corn

I concede the limit. I retain the preference.



Herman

Fusion's experiments suggest that some remaining audio weakness lies inside the frozen language model's ability to process audio tokens. A better connector can't force an unchanged network to develop every useful acoustic distinction. That's why the deeper adapters matter.



Daniel

If I take your local recommendation, what is the actual setup? Not commands, just what has to be in place.



Corn

Start with the retrieval checkpoint, not the clustering variant. Jina omni nano is roughly zero point nine five billion parameters and produces seven hundred sixty-eight dimensional vectors. Small enough to make a local trial plausible, but nano does not mean no memory planning.



Herman

Its maximum sequence length is eight thousand one hundred ninety-two tokens. Install the Transformers and PyTorch stack, plus Pillow and NumPy. The documented minimums are Transformers four point five seven and PyTorch two point five. Then add the readers for the modalities you actually intend to ingest.



Corn

Audio needs decoding support such as librosa and soundfile. Video needs something such as torchcodec or av. PDFs need page handling libraries such as pdf two image or pypdfium two. Installing the model doesn't magically install every media format's problems out of existence.



Herman

For serving, the documented vLLM path has been validated on version zero point twenty point one. I'd pin a working environment rather than assemble the newest version of everything and call the resulting argument progress.



Corn

You've become very wise since your last dependency upgrade.



Herman

I can learn from an afternoon, Corn.



Daniel

Once it runs, am I just encoding everything and putting the vectors in my database?



Herman

Almost... no, that misses an important instruction. Retrieval distinguishes the query side from the document side. Use the explicit query and document encoding methods, or the documented task prompts. On the raw path, those can be query and document prefixes. A bare generic encoding call doesn't select the required retrieval behavior.



Corn

Keep the original source and metadata alongside the vector. A video result still needs a timestamp. A PDF result still needs a page. One shared embedding space doesn't relieve you of telling the user where the answer came from.



Herman

Mixed content can go through one call, such as a text description with an image. Decide whether you need a fused item or separately searchable pieces. That's a product decision before it's a batching decision.



Daniel

I can imagine a whole video embedding returning the right video but not the right moment.



Corn

Precisely the problem. Segment it at useful boundaries, retain the time ranges, and test both finding the item and locating the passage. A correct title with twenty minutes of hunting afterwards is only partly a search result.



Herman

One implementation trap deserves special attention. The Whisper feature extractor pads clips to a thirty second window. A naive token count can emit seven hundred fifty audio tokens for a two second clip.



Corn

The silence has excellent representation.



Herman

Size the audio run using the real, unpadded frame mask. Otherwise you're presenting the model with a token pattern unlike its training. That is a preprocessing error, not proof that omni embeddings dislike your recordings.



Daniel

Where does that leave Gemini as the managed option? Is the larger output vector itself an advantage?



Herman

Gemini's full vector has three thousand seventy-two dimensions, with Matryoshka training optimized for reductions to seven hundred sixty-eight and fifteen hundred thirty-six. You can trade vector storage and search cost against retained information. More coordinates don't guarantee better relevance, though.



Corn

Nor does less setup guarantee a lower bill. We don't have a verified price comparison here. I'd compare real ingestion costs, media limits and search quality before choosing the API route.



Herman

Gemini reports strong results, including sixty-two point nine recall at one on MSCOCO image retrieval. That means the correct result must occupy the first position. Its claim to surpass specialists is meaningful counterevidence to the idea that every generalist is inevitably worse.



Daniel

But sound is the place you would be most suspicious?



Herman

For a retrofit, yes. Vision and language backbones have become strong at text, images and video, but audio isn't simply another image format. Acoustic information has to survive its encoder, the connector and the receiving model.



Corn

Your DJ ear would care whether a recording contains a cymbal or a leaking pipe. A text description might call both a hiss. The model has to retain the distinction before a query can recover it.



Herman

Would you use one system for a mixed archive even if its sound search was noticeably weaker, Corn?



Corn

If audio were a small convenience feature, possibly. If sound were the archive's reason to exist, no. Saving an integration doesn't compensate for failing the thing people came to find.



Herman

Fusion gives us a concrete comparison. On AudioCaps audio to text retrieval, its first generation puts the correct caption in the top ten about seventy-four percent of the time. The second generation reaches about seventy-eight percent. CLAP family specialists are around ninety-one to ninety-three percent on that evaluation.



Corn

That gap would be a downer if I were building sound search. Mum has apparently found a more domestic way to deliver it.

D

Dorothy

Hello dear, it's your mother. Time for my grim fact.

D

Dorothy

This one's about the audio, sweetheart. The clever general models score nought point seven four one on finding the right sound, and the old specialists get nought point nine two. So all that lovely audio, all those recordings nobody specialized in, just sits there getting a worse answer than it used to.

D

Dorothy

Anyway, are you eating? I've got a casserole and it won't keep forever.

D

Dorothy

Right, I'll let you go. Bye bye, dear.



Herman

Lovely to hear you, Mum.



Daniel

Does that comparison establish that I should always keep an audio specialist?



Herman

It establishes a substantial gap for that system on that task. It doesn't establish a permanent tax on every generalist. Gemini claims specialist beating results on its chosen benchmarks, and native omni backbones have more freedom to learn acoustic distinctions.



Corn

Also, audio to text caption retrieval isn't every sound problem. Finding a caption, finding a similar noise, and finding an image associated with a recording ask different things of the space.



Herman

Fusion manages audio to image retrieval without audio visual training pairs, reaching roughly forty-two percent recall in the top ten on its VGGSound evaluation. That's around twenty-nine times chance. Weakness relative to an audio specialist can coexist with a useful cross-modal capability that specialist doesn't offer.



Daniel

That sounds like the real attraction. A specialist might win its test but not let me search across the rest of the archive at all.



Corn

You can use specialists together, but then you own the routing and score combination. Their similarity scores aren't automatically comparable. The image system's confident match and the audio system's confident match need not mean the same probability of relevance.



Herman

A shared space reduces that integration burden. Still, don't assume a sound query and a text query produce equally calibrated rankings merely because they're supported. I'd test modality specific slices alongside the mixed results.



Daniel

What about documents? I don't want PDF support to mean somebody quietly took a screenshot and discarded the useful structure.



Corn

Direct file acceptance and document understanding are separate promises. Jina's card explicitly lists PDFs. That makes it the clearest documented local path here. It doesn't establish a distinct layout aware modality with guarantees about tables, reading order or tiny footnotes.



Herman

The other systems can represent document content through images or interleaved text and images. In this set, I haven't seen a separately defined layout input that establishes superiority on document retrieval. Accepting a PDF tells you about ingestion. Finding the right table cell requires its own test.



Corn

The extension has passed. The footnote remains under review.



Herman

A scanned invoice and a digitally generated manual can take very different processing paths. I'd preserve page references and inspect what the model actually receives, especially when rendering determines whether small text survives.



Daniel

I'm hearing that there isn't a useful combined leaderboard. How do I avoid spending a month comparing incompatible claims?



Herman

There isn't an independent apples to apples comparison I can point you to across all the requested inputs. Gemini, Ovis and Jina report different benchmark mixes. Even shared benchmark names can conceal different training exposure or evaluation settings.



Corn

Pick representative queries from the job you want done. Include text finding images, sound finding captions, and documents finding passages. Keep a separate audio slice so a large pile of easy text queries can't conceal poor acoustic retrieval.



Herman

Training exposure matters too. M two D CLAP includes Clotho in training, so its Clotho result isn't zero shot. That doesn't make it useless. It changes what the score demonstrates about transfer to unfamiliar material.



Daniel

Would you test the defaults first, or start adjusting dimensions and preprocessing immediately?



Corn

The documented retrieval defaults first. Otherwise you won't know whether you evaluated the model or your improvements to it.



Herman

Some apparent improvements make things worse. Fusion's cleaner captions, rewritten by a language model, reduced retrieval. A supposedly better audio tower lost sixteen recall points inside the frozen model splice. Wider connectors could memorize rather than generalize.



Corn

The upgrade needed a downgrade.



Herman

I suspect some rewritten captions remove rough but useful acoustic specificity... that's an interpretation, not a demonstrated explanation for every case. The established result is that polished text didn't improve that retrieval system. Likewise, a stronger standalone audio encoder needn't be a better source of features for a frozen recipient.



Daniel

Then your recommendation is a controlled comparison rather than trying to build the perfect pipeline first?



Corn

A small comparison with the intended preprocessing. Keep the candidate shortlist short enough that you actually inspect the misses. That will tell you more than collecting every score ending in state of the art.



Herman

In particular, inspect the near misses. A recording of rainfall finding running water may reveal useful acoustic structure. Finding a caption that merely shares the word outside reveals something else. The error tells you what additional training or a specialist might buy.



Hilbert

Herman, does your query side need to know whether the seed packet is damp? Nissim at the seed shop is asking me to separate those before he will continue scanning. I'm building a search index for his packets this month. Photographs, printed descriptions and cassette notes. The shop is still taking orders while I lick the packets for the flavor coordinate.



Corn

Put dampness in the metadata. You'll want to filter it without rebuilding the vectors.



Hilbert

That's what I tell him. My cousin handles his catalogue and says a single number is enough. Color, seed count, acquisition date. He calls it one vector space. I don't trust his opinion. He also calls a photograph of the back a second variety. His view is that audio adds nothing because everything useful is already printed on the packet. Then he dictates corrections into my recorder.



Herman

That would make the recordings useful even before you consider the sound itself.



Hilbert

They are useful. That's the bit I like, actually. A spoken correction finding the printed description without anyone typing it again. Corn, your point about returning the page is making me insist on the packet number. Nissim keeps sending me a photograph of the shelf instead. As an anteater I have to turn sideways to read his handheld scanner, and he charges the scanning by the minute.



Corn

The shelf photograph makes the customer finish the search.



Hilbert

I'm already paying him to hold a tuning fork beside each packet while he records its name. He says the catalogue needs a reference tone. The cassette notes now contain more tuning fork than description. This morning he invoices eleven minutes for radishes.



Herman

Eleven minutes for radishes is enough to make ingestion cost part of the model comparison. I'd like an independent evaluation that includes that whole path, from original media to a useful located result. Until then, I'd trial Jina omni for local, index compatible mixed retrieval and Gemini for the managed native comparison, while keeping an audio specialist in contention for audio heavy work.



Corn

Thanks to Hilbert Flumingtop for producing.



Herman

For more along these lines, there's episode five, Fine-Tuning ASR For Maximal Usability; episode eleven, How Does Fine Tuning Work Anyway; and episode four, If Your Voice Ages, Does Your Fine-Tune Become Useless. This has been My Weird Prompts, the human AI collaboration podcast.



Corn

You can find the show at my weird prompts dot com.



Daniel

See you soon.



Corn

Goodbye.

SOURCES

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

1. **jina-embeddings-v5-omni / GELATO technical report, 2026-05 [primary]**
https://arxiv.org/html/2605.08384

2. **Gemini Embedding 2: A Native Multimodal Embedding Model from Gemini, 2026-05 [primary]**
https://arxiv.org/html/2605.27295

3. **model card, setup instructions, supported file types [primary]**
https://huggingface.co/jinaai/jina-embeddings-v5-omni-nano-retrieval

4. **model card, clustering variant [primary]**
https://huggingface.co/jinaai/jina-embeddings-v5-omni-nano-clustering

5. **repo, One model. One vector space. Text, image, video, audio. [primary]**
https://github.com/Eximius-Labs/fusion-embedding

6. **Elastic press release, 2026-05-11 [primary]**
https://ir.elastic.co/News--Events/news/news-details/2026/Elastic-Introduces-Jina-v5-Omni-Family-Two-Models-to-Power-Text-Image-Video-and-Audio-Search/default.aspx

7. **e5-omni-7B model card [primary]**
https://huggingface.co/Haon-Chen/e5-omni-7B

8. **Fusion Embedding: A Unified Embedding Space for Text, Image, Video, and Audio, 2026-07**
https://arxiv.org/html/2607.18666

9. **Ovis-Embedding: Pushing the Frontiers of Universal Omni-Modal Embeddings, 2026-09**
https://arxiv.org/abs/2609.25165

10. **Elastic Search Labs, 2026-08-12**
https://www.elastic.co/search-labs/blog/multimodal-embeddings-gelato-jina-v5-omni

11. **TSEmbed: Unlocking Task Scaling in Universal Multimodal Embeddings**
https://arxiv.org/html/2603.04772v1

12. **jina-embeddings-v4, ACL Anthology**
https://aclanthology.org/2025.mrl-main.36/

13. **GME: Improving Universal Multimodal Retrieval by Multimodal LLMs**
https://arxiv.org/html/2412.16855

14. **WAVE: Learning Unified & Versatile Audio-Visual Embeddings with Multimodal LLM, ICLR 2026**
https://arxiv.org/pdf/2509.21990

15. **Omni-Embed-Nemotron: A Unified Multimodal Retrieval Model for Text, Image, Audio, and Video**
https://arxiv.org/html/2510.03458

16. **Tianmu-Emb-Uni-8B model card**
https://huggingface.co/TianmuLab/Tianmu-Emb-Uni