Episode Book

EPISODE #6311

How Transformers Process Images and Video

PUBLISHED	27 October 2026 at 19:06 UTC
DURATION	29:48
LISTEN	myweirdprompts.com/episode/transformers-vision-diffusion-patches-24706/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.22
SOURCES	19 cited by the research (listed at the end)
BOOK GENERATED	28 October 2026 at 09:53 UTC

SYNOPSIS

Transformers originally built for language now power visual perception and image generation. Replacing convolutional networks with self-attention requires turning continuous pixels into discrete vector sequences, which Vision Transformers achieve by treating raw image patches like text tokens. Exploring architectures like ViT and Diffusion Transformers (DiT) shows how swapping inductive visual biases for massive pretraining datasets allows attention mechanisms to handle spatial relationships, denoising targets, and latent decoding without treating pictures like sentences.

TRANSCRIPT



Corn

I find it faintly insulting that a poetry engine can also draw better than me.



Corn

Daniel is asking how the transformer stretches across text, pictures, speech, transcription, and moving scenes without becoming a completely different machine. Those are different kinds of information, with different demands on what counts as getting something right. A sentence can survive a changed adjective. A voice can stop sounding like its owner while every word remains correct.



Herman

A video can preserve every object and still have the cup pass straight through the table.



Corn

Daniel wants to know what gets rebuilt when the architecture crosses those boundaries, and what actually survives the journey. The units it sees, the thing it predicts, the way an answer emerges. I suspect you like this because it sounds like replacing every appliance in the flat with one extremely expensive box.



Herman

Before you object to the box, Corn, the important distinction is between sharing a design and sharing trained knowledge. A vision transformer and a language transformer can use closely related machinery without having the same weights or understanding each other's inputs.



Corn

Your universal appliance needs separate plumbing.



Herman

Quite a lot of it. The short answer is that the transformer barely changes. The long answer is everything else.



Herman

The original transformer, introduced by Vaswani and colleagues at Google in two thousand seventeen, takes vectors and repeatedly mixes information between them. Attention computes how strongly one position should draw on other positions. Multiple heads can learn different relationships, and feedforward layers transform what each position now contains. Residual connections keep information moving through the stack.



Corn

The word vector matters here. People hear token and imagine a numbered word from a dictionary. The transformer doesn't require that. Give it a projected image patch, and the arithmetic still works. It doesn't complain that the noun has cheekbones.



Herman

Right... well, there's a correction worth making before we go further. Bare self-attention is permutation equivariant, not permutation invariant. Shuffle the input vectors and its output vectors shuffle correspondingly. It doesn't inherently know which arrangement means something.



Corn

It can preserve the seating plan without knowing where anyone belongs.



Herman

Positions supply that missing structure. Text needs order along a sentence. Images need horizontal and vertical location. Video needs time too. Those coordinates can enter through learned embeddings, other positional encodings, or the attention calculation itself. A causal mask supplies another constraint, preventing a text generator from inspecting words it hasn't generated yet.



Corn

Which means the supposedly universal part still gets told where it is and what it's allowed to see.



Herman

Those choices matter enormously. We're preserving a family of computations, not promising identical wiring in every implementation. What makes it general is that the layer stack imposes remarkably little about the meaning of each vector.



Corn

I'd call that a universal interface rather than a universal understanding. There is a lot of engineering hidden in the word input.



Herman

The transformer is the easy part, relatively speaking. Turning a domain into useful vectors, choosing the learning problem, and turning predictions back into something usable often carry the hardest work. Vision transformers make that visible at the input. Diffusion transformers expose it at the objective. Audio codecs and video compression make the costs impossible to ignore.



Corn

You're going to fall for the neatness of this, Herman.



Herman

Already have. Now we get to damage the neatness.



Corn

Images give us a particularly honest test. A convolutional network arrives with strong opinions about pictures. Nearby pixels probably belong together. An edge remains an edge if it moves across the frame. A plain transformer doesn't arrive with those opinions built in to the same degree.



Herman

ViT takes an image, cuts it into patches, flattens each patch, and applies a learned linear projection. That produces the vectors. The original paper says, “Image patches are treated the same way as tokens” in a language application. That's a startlingly small conceptual move.



Corn

Sixteen by sixteen pixels becomes a token. Not a cat token, not an ear token. Just a square that may contain half an ear and some wallpaper.



Herman

The network learns which combinations matter. Consider an image two hundred twenty-four pixels across, split into those patches. You get a grid fourteen patches across, and fewer than two hundred patch tokens overall. That is manageable for full attention.



Corn

Wallpaper finally gets equal representation.



Herman

Pixel tokens would be vastly more expensive. Dense attention forms relationships between pairs of positions. Double the number of positions and the attention matrix grows roughly fourfold. Starting with patches reduces the sequence before that multiplication becomes painful.



Corn

There's a small trap in calling that compression. A linear projection needn't discard information if the dimensions permit it. We're definitely reducing the number of positions the attention layer handles, but we aren't necessarily doing the same kind of lossy compression as a video codec.



Herman

Fair correction. Sequence compression, not automatically information compression. The patch can retain internal detail in its vector. Attention then works between patches, while later transformations learn how those internal details should contribute.



Herman

Would you choose that over a convolutional network if you had a modest collection of labelled pictures, Corn?



Corn

Probably not. I'd want the machine to arrive knowing something about pictures rather than charge me to rediscover adjacency. If my data budget is small, useful assumptions are a feature.



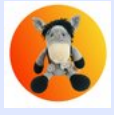
Herman

That was the original difficulty. On the ordinary ImageNet training scale, ViT could fall a few points behind comparable ResNets. Remove locality and translation-related assumptions, and you need enough examples to learn relationships those assumptions supplied cheaply.



Corn

The machine requires three hundred million examples of nearby things.



Herman

Large pretraining collections changed the outcome. With the huge JFT collection, the reported ImageNet accuracy reached eighty-eight point five five percent. The lesson wasn't that visual structure had become irrelevant. Data and model capacity could compensate for not hardwiring as much of it.



Corn

That distinction matters when somebody turns the result into “architecture no longer matters.” They bought flexibility with data. If I can't afford the data, I don't get the same bargain.



Herman

Or if your images differ sharply from pretraining. Medical scans, industrial defects, unusual sensors. Transfer can help enormously, but the match between representation and task remains important. A general backbone doesn't abolish the question of what examples it has learned from.



Corn

You're resisting the opportunity to prescribe a larger model.



Herman

I'm a retired pediatrician. Occasionally I still check the patient.



Corn

There he is.



Herman

The patch size itself makes a practical choice. Smaller patches expose finer spatial structure to attention but lengthen the sequence. Larger patches shorten it but demand more work inside each representation. A tiny feature can matter even when it occupies only a fraction of a patch.



Corn

Especially if you're looking for a defect rather than deciding whether the whole picture contains a bicycle. “Mostly intact” is not a useful verdict on a brake cable.



Herman

Classification was a convenient demonstration because a pooled representation or special classification token can produce the answer. Dense prediction asks for more. Segmentation needs spatially resolved outputs. You may need additional decoding layers that recover a detailed map from the transformer's representations.



Corn

Already the same trunk has different exits. That helps me separate the claim that survives from the sales pitch. The stack can transfer. The entire system doesn't become taskless.



Herman

Generation makes that separation even clearer. DiT keeps latent diffusion but replaces the convolutional U-Net doing the denoising with a transformer. The image first enters a learned latent representation. Patches of that latent become the transformer's inputs.



Corn

We're not cutting the photograph into little vocabulary entries and asking for the next square?



Herman

No. In the original DiT setup, the model receives a noisy latent, the diffusion timestep, and conditioning such as an image class. It predicts noise. A sampling procedure uses those predictions to progressively remove noise. Finally, a learned decoder turns the resulting latent into pixels.



Corn

The attention stack supplies an estimate. Something outside it runs the repeated clean-up.



Herman

Yes, though “outside” means the surrounding generation procedure, not necessarily a separate service. It's worth separating the denoising network, the numerical sampler, and the latent decoder. People call all three the decoder and then wonder why the explanation feels slippery.



Corn

Three things wearing one name. Very efficient documentation.



Herman

DiT's headline result was an FID of two point two seven on class-conditioned ImageNet generation. Lower is better for that metric. At the time, that was a leading result, achieved after replacing a backbone people associated strongly with successful diffusion.



Corn

FID isn't a percentage of pictures that look right, though. It compares distributions of image features. It can reward a model without answering whether my requested bicycle has working pedals.



Herman

Nor does a class-conditioned benchmark establish excellent instruction following. The significance here is narrower and cleaner. A transformer could do the denoising job extremely well. Convolutional U-Net structure wasn't indispensable.



Corn

I can see a confusion forming. We describe text generation as predicting something missing and image diffusion as predicting something missing. They're still very different training tasks.



Herman

Very. Text commonly uses a categorical distribution over a vocabulary, trained against the actual next token. Diffusion learns a continuous prediction associated with a noise process. Both backpropagate through attention layers, but the output heads, losses, conditioning, and sampling schedules differ.



Herman

Do you think sharing the layers is meaningful enough to justify calling it the same architecture?



Corn

Yes, provided we don't sneak “same model” into the next sentence. An engine design can be reusable without the entire vehicle being interchangeable. What interests me is that attention can discover useful interactions without being born with a theory of pixels. That is substantial. Claiming images are secretly sentences adds nothing.



Herman

I agree with the distinction. Also... I almost said the trunk is untouched. Nearly untouched is better. DiT uses conditioning through adaptive normalization, among other design choices. The family resemblance is strong, but implementations aren't photocopies.



Corn

You were enjoying that tidy diagram.



Herman

I can enjoy it with annotations.



Corn

Your annotations need their own diagram.



Herman

The result changes how researchers search. Instead of inventing a wholly separate interaction mechanism for images, they can reuse optimization experience, scaling practices, and transformer infrastructure. Then they spend ingenuity on representations and learning dynamics.



Corn

That includes a dependency people tend to hide. If the latent encoder discards fine lettering, the transformer can't recover the exact lettering merely by being enormous. It may invent something plausible instead.



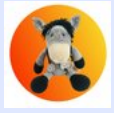
Herman

A bottleneck upstream sets a ceiling. Better generation can conceal that ceiling aesthetically, which is dangerous if the thing you need preserved is precise. Reconstruction tests, instruction tests, and perceptual quality ask different questions.



Corn

My favourite beautiful picture is the one containing the correct invoice total.



Herman

You'd be unpopular on an image-model launch team.



Corn

I'd read the receipt.



Herman

There's another consequence of diffusion. Image positions can attend bidirectionally during denoising. The upper left can use the lower right while both are being refined. Nothing about transformer arithmetic requires every application to generate strictly from left to right.



Corn

That seems central to Daniel's question. The portable mechanism is attention over representations, not the particular habit of finishing one word and committing forever.



Herman

Text made that habit familiar. Images expose how optional it is. Audio will make the choice more complicated still, because some applications can inspect the future and others have someone waiting on the telephone.



Corn

How does that one idea, attention over sequences learned at scale, stretch across all of them without forcing every output into the same production line?
Ivan: Lighthouse keeper, lighthouse keeper. Seventeen twenty nine. Forty four eleven. Seventeen twenty nine. The kettle is not on the third stair. The curtain in the spare room has been counted twice. The tokenizer sleeps in the garden. Seventeen twenty nine. The lamp is swept, the lamp is swept. Nightjar.



Herman

Wrong number.



Corn

What on earth just came through the desk?



Herman

Speech makes the production line depend on latency. A live conversation can't wait for a complete future utterance, whereas editing a recording can use material on both sides of a gap.



Corn

Audio also makes the token question harder. A patch of wallpaper doesn't unfold at thousands of samples per second. Feed raw waveform samples into ordinary full attention and the sequence becomes absurd before anyone finishes saying hello.



Herman

Neural audio codecs learn a compact representation, then quantize it into discrete codes. A language-style model can predict those codes. A codec decoder reconstructs the waveform. Several codebooks may capture successive levels of detail, so one moment of sound can require multiple symbols.



Corn

The compression needs to preserve more than words. A transcription system might profit from ignoring my microphone. A voice generator needs enough information to reproduce breath, pitch, rhythm, and the difference between speaking and singing.



Herman

That tension is why “best codec” is an incomplete phrase. Better reconstruction doesn't guarantee better downstream generation. A codec can preserve intricate acoustic details while producing a code sequence that is harder for the generative model to predict.



Corn

It saves the noise beautifully and then makes the composer learn it.



Herman

Or spreads meaningful distinctions across many codes in an inconvenient way. You need to evaluate the codec with the model that will use it, not just listen to reconstructed recordings.



Herman

Corn, if a codec made your voice slightly less exact but halved the conversational delay, would you take it?



Corn

For a call, yes. For an audiobook, probably not. I'd also want to know whether “slightly less exact” means flatter emotion or a different consonant. One is disappointing. The other changes what I said.



Herman

That divides the objectives neatly. Intelligibility, speaker identity, naturalness, latency, and acoustic fidelity can move separately. A single reconstruction score won't tell you which compromise you've made.



Corn

Also, a thousand tokens of audio don't buy the same amount of meaning as a thousand tokens of text. Context length on the product page isn't a unit of conversation.



Herman

Codec sequences can be very long. That pressures memory, attention, and autoregressive decoding time. More aggressive compression helps length, but can make each remaining token carry a more difficult prediction.



Corn

The bill moves between departments.



Herman

Interestingly, codec-token sequences show language-like frequency patterns, particularly in groups of three tokens. Common combinations recur often, and vocabulary growth has familiar statistical shapes. That offers useful tools for analysis.



Corn

It doesn't mean the third token is a verb.



Herman

No. Similar statistics don't make acoustics equivalent to grammar. They show structure a predictor might exploit. The codes can reflect articulation, recording conditions, or speaker characteristics as well as linguistic content.



Corn

We should keep the continuous route visible too. Otherwise we've just told everyone that sound has to become discrete before attention can touch it.



Herman

It doesn't. Voicebox works with continuous speech features and flow matching, trained around text-guided infilling. It can generate a missing region using context on both sides rather than predicting every codec symbol sequentially.



Corn

The future context is available because the recording already exists. For a live response, it can't borrow tomorrow's syllables.



Herman

It can still generate a planned chunk non-autoregressively, but streaming adds constraints. In the reported zero-shot comparison with VALL-E, word error rate fell from five point nine percent to one point nine percent. That means fewer transcription errors in the generated speech under that evaluation.



Corn

A useful improvement, not proof the voice sounds human.



Herman

Speaker similarity improved separately, and the reported speedup reached twentyfold in some settings. Those are distinct outcomes. Flow matching changes the training target and generation dynamics while keeping transformer-style interactions.



Corn

Transcription points in the opposite direction. Whisper doesn't need to reconstruct my wheeze. It encodes audio features, then a text decoder attends to those features while producing words.



Herman

Cross-attention lets the output sequence consult a different input sequence. Their lengths don't have to match. Audio frames and written tokens get aligned through learning, without requiring one audio token for every word.



Corn

That is a more convincing universality than pretending the units match. The same machinery can connect representations that remain different.



Herman

Video then adds a problem that a transcript can avoid. You need visual detail and motion over time. Sora's original description uses spacetime latent patches. A compression model maps video into a latent space, and patches represent small regions extending across space and time.



Corn

A patch gets a duration. More room for a cup to commit a crime against furniture.



Herman

The denoising model must coordinate those patches. Appearance has to persist, motion has to connect, and objects shouldn't become unrelated objects between frames. Compression reduces the burden, but long clips still produce enormous sequences.



Corn

Increase the spatial resolution and you add patches across both image dimensions. Increase duration and you add another dimension of positions. Then dense attention compares pairs across the result. That isn't ordinary feature creep.



Herman

It's why sparse and windowed attention return. Nearby space and nearby time contain strong correlations. Different layers or heads can emphasize local relations while selected routes preserve longer-range communication.



Corn

Vision transformers made headlines by learning around missing locality assumptions. Video puts locality back on the purchase order.



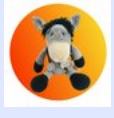
Herman

Scale could pay for learning flexible relationships. It doesn't make every pairwise interaction worth computing. Compact Attention reports attention speedups of roughly one point six to two point five times using structured spatial and temporal sparsity.



Corn

Attention speedups, not necessarily the whole video arriving twice as fast.



Herman

Correct. Encoders, decoders, other network operations, and the sampling loop still take time. Saving time in one component doesn't multiply identically through the complete system.



Corn

I also wouldn't call those attention changes trivial. We're keeping the family, but choosing who can talk to whom is architecture.



Herman

Agreed. “Nearly unchanged trunk” should never mean “all attention patterns are equivalent.” Video is precisely where domain structure can reshape the computational pattern even when the basic layer operations remain recognizable.



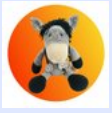
Herman

Now put text and images into one model. Would you force both through discrete tokens, Corn, or let them retain different prediction rules?



Corn

I'd keep different rules until somebody showed the shared rule was worth its cost. Quantizing a picture just to preserve the elegance of next-token prediction feels like making a photograph pass a spelling test.



Herman

Transfusion takes the hybrid route. One transformer predicts text tokens and denoises continuous image representations. Text uses causal attention. Image patches can attend bidirectionally within the image.



Corn

That is a single network with different working conditions inside the sequence.



Herman

In its ablations, forcing causal attention within the image hurt image generation. So even when weights are shared, the access rules should reflect the task.



Corn

The most useful common part survives because they leave the differences alone.



Herman

Compared with the discrete-image Chameleon approach, Transfusion achieved roughly half the FID at matched compute in the reported comparison. Its image-to-text results also reached parity using much less compute. That supports the hybrid choice in that setup, rather than proving continuous representations win everywhere.



Corn

Compression contributes too. The modality-specific layers can represent an image with just sixteen patches in one configuration. But sixteen vectors are not sixteen words' worth of information.



Herman

Nor is the reported potential serving-cost reduction a universal multiplier. The chosen image representation changes how much work the transformer must do. Specialized boundary layers can shoulder work that otherwise lands in a long sequence.



Corn

We're calling the center universal while the edges do increasingly specialized labour.



Herman

There's still no agreed universal tokenizer. AVTok targets a shared audio-video representation, but that pair doesn't settle text, images, sound, and everything else together.



Corn

I wouldn't assume one alphabet is the destination. Even humans preserve diagrams because describing every arrow in prose is dreadful.



Herman

Sharing parameters creates another complication. Understanding and generation can push those parameters in conflicting directions. A recent position-resolved analysis measured negative alignment between their gradients, strongest near the start of the visual sequence.



Corn

Meaning one task's proposed update can locally make the other's job worse. Not merely “the model has lots to learn.”



Herman

Yes. The early visual positions had a mean gradient cosine around negative zero point one eight. That's an optimization measurement in that setup, not a fixed tax on every multimodal model. It gives a concrete mechanism for interference.



Corn

Training balance then matters as much as input balance. Half the examples being images doesn't guarantee half the learning benefits images.



Herman

Vision can also be more data-hungry than language. Matching their token counts doesn't equalize their progress. Loss weighting, staged training, and separate components become practical choices rather than blemishes to hide.



Corn

The MMMG evaluation is a useful brake on the idea that combining capabilities guarantees mastery. Its best image score was seventy point seven percent, with audio and interleaved tasks doing worse. Those are task scores, not the fraction of all possible images a model understands.



Herman

Specialists can still lead. A shared backbone gives you a place to combine capabilities, not an automatic improvement in every one.



Corn

I suspect this is why people talk past one another about tokenizers. A codec code, a projected patch, and a text token all occupy a sequence position, but they promise very different things.



Herman

Even the text case has edge cases. An isolated space is always just one token, for instance, while a word may split across several. The boundary rules are part of the system you're committing to maintain.



Hilbert

Not always one, Herman. The counter at Barak Office Systems shows two for the blank field in their quotation tool. I'm disputing it with Nissim this month. I go there for a document-scanning appointment, and their new assistant keeps inserting the token “ceremonial” before “replacement roller.” The tokenizer accepts the word perfectly. Nissim says that proves it isn't a bug. He is charging me for a ceremonial replacement roller consultation every Thursday.



Corn

The ceremony appears to be weekly billing.



Hilbert

The consultation is compulsory because I'm now the shop's nominated witness for that token. Nissim has my chair labelled with it. I don't sit comfortably because the armrest catches my anteater claws. The label is beside my name, not instead of it. I insist on that. Their maintenance contractor says the tokenizer is behaving. The assistant is choosing the wrong valid token. That's the distinction, Herman, after your very careful separation of the sampler from the decoder.



Herman

A valid vocabulary entry tells them the word can be represented. It doesn't establish that the generated description is correct.



Hilbert

My cousin Pavel maintains these systems. He says changing the tokenizer would be the wrong repair. Then every saved conversation needs checking, the quotation templates need checking, and somebody needs to answer the telephone when a space moves. He says keep the vocabulary and correct the examples or the generation. He is not to be trusted. He invoices Nissim by the unresolved adjective. At present, “ceremonial” is his only regular income. Nissim offers me an ordinary roller appointment if I withdraw the complaint, but the appointment form still includes the word. The actual maintenance window is eleven minutes. The argument takes the rest of Thursday.



Corn

They could leave the token alone and reject that description before a quotation reaches you.



Herman

A constrained output field would make more sense than rebuilding the vocabulary.



Hilbert

Yes. The field still has to say roller.



Corn

Eleven minutes of maintenance, surrounded by a dispute about which component owns the mistake. That's an interface cost no parameter count is going to advertise.



Herman

One adjacent detail we haven't used is joint audio-video generation with separate noise schedules for the two modalities. Even inside one system, sound and pictures needn't progress through generation at identical rates.



Corn

Synchronizing the result doesn't require pretending the ingredients behave alike.



Herman

It requires coordinating them.



Corn

The development I'll watch is whether shared models make those boundaries easier to maintain, rather than simply hiding them. Attention can keep absorbing new kinds of input, but someone still has to decide what each representation preserves, how it learns, and which component fixes an incorrect output.



Herman

Our thanks to Hilbert Flumingtop for producing the show.



Corn

If this was your kind of episode, go back for 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 these conversations, leave us a review on your podcast app. It helps other curious people find the show.



Corn

See you tomorrow.

SOURCES

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

1. **Dosovitskiy et al., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, arXiv:2010.11929 [primary]**
   https://arxiv.org/abs/2010.11929
2. **Peebles & Xie, Scalable Diffusion Models with Transformers (DiT), arXiv:2212.09748 [primary]**
   https://arxiv.org/abs/2212.09748
3. **Le Lan et al., Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale, arXiv:2306.15687 [primary]**
   https://arxiv.org/abs/2306.15687
4. **Zhou et al., Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model, arXiv:2408.11039 [primary]**
   https://arxiv.org/abs/2408.11039
5. **Liu et al., Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models, arXiv:2402.17177 (Feb 2024)**
   https://arxiv.org/abs/2402.17177
6. **Park et al., Analysing the Language of Neural Audio Codecs, arXiv:2509.01390, ASRU 2025**
   https://arxiv.org/abs/2509.01390
7. **Li et al., Investigating Neural Audio Codecs for Speech Language Model-Based Speech Generation, arXiv:2409.04016, SLT-2024**
   https://arxiv.org/abs/2409.04016
8. **Liu et al., Speech Token Prediction via Compressed-to-fine Language Modeling, arXiv:2505.24496**
   https://arxiv.org/abs/2505.24496
9. **Miao et al., Kaleido: Algorithm-Hardware Co-Design for Video Diffusion Transformers, arXiv:2607.13770 (2026-07-15)**
   https://arxiv.org/abs/2607.13770
10. **Li et al., Compact Attention: Exploiting Structured Spatio-Temporal Sparsity for Fast Video Generation, arXiv:2508.12969**
    https://arxiv.org/abs/2508.12969
11. **Jiang et al., Beyond Layers: Position-Resolved Gradient Conflict..., arXiv:2609.38485 (2026-09-29)**
    https://arxiv.org/abs/2609.38485
12. **Tong et al., Beyond Language Modeling: An Exploration of Multimodal Pretraining, arXiv:2603.03276 (2026-03-03)**
    https://arxiv.org/abs/2603.03276

13. **Wikipedia, Transformer (deep learning architecture), last edited 2026-10-26**
(deep_learning)
https://en.wikipedia.org/wiki/Transformer

14. **The Revision, New Benchmark Exposes Gaps In AI Image and Audio Generation, 2026-10-02**
https://therevision.co/articles/new-benchmark-exposes-gaps-in-ai-image-and-audio-generation

15. **VideoTech Wire, UniAVC Targets More Flexible Audio-Video Creation and Editing, 2026-10-20**
https://videotechwire.com/news/uniavc-unified-audio-video-creation-via-progressive-training-and-asymmetric-scheduling/

16. **RuntimeWire, MiniMax brings synchronized audio-video generation to Reactor's API, 2026-10-13**
https://runtimewire.com/article/minimax-fasth3-omni-synchronized-audio-reactor

17. **Google, EmbeddingGemma 2, 2026-10-06**
https://blog.google/innovation-and-ai/technology/developers-tools/embeddinggemma-2/

18. **ai.info, Apple Details a 3B Tri-Modal Diffusion Model Trained on 6.4T Tokens**
https://ai.info/news/apple-tri-modal-masked-diffusion-study

19. **Banandre, MiniMax H3 Is Here**
https://www.banandre.com/blog/minimax-h3-open-source-video-model-audio-generator