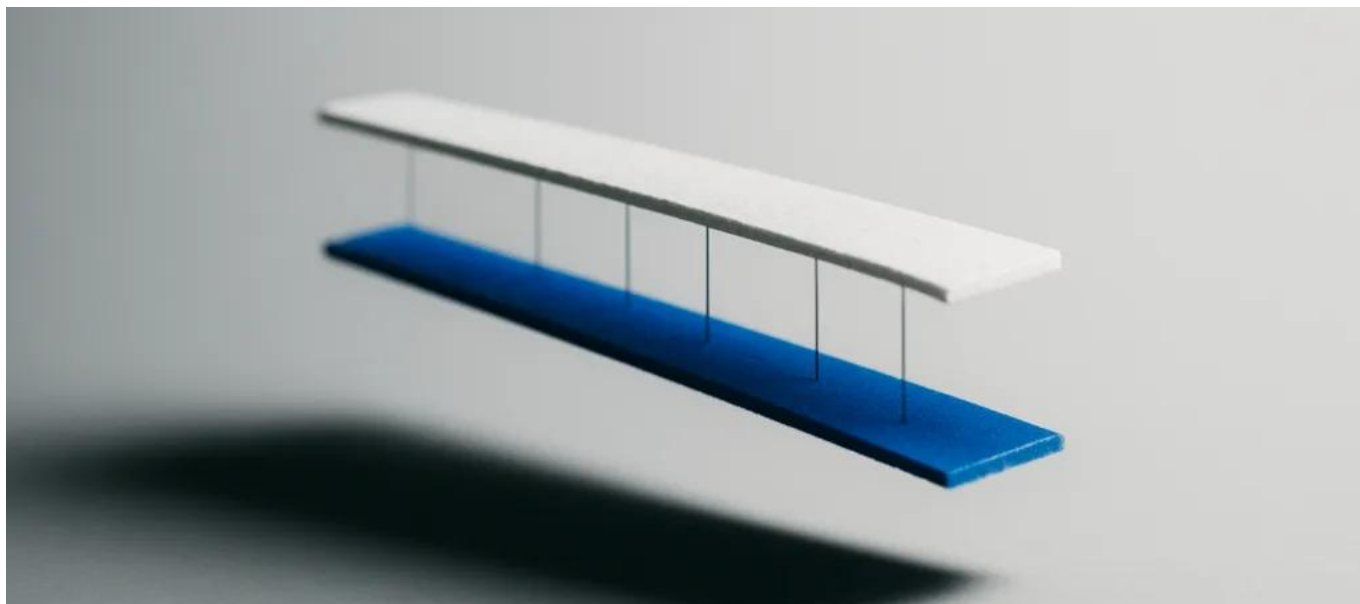




CrossDiT, Explained: How Kandinsky 6.0 Generates Video and 44 kHz Audio in One Pass, and Where It Breaks

Kandinsky Lab shipped a 29B audio–video model under MIT on 6 October 2026, weights and code included. The clips it produces are five seconds long.

Both facts are true at once. CrossDiT, the architecture described in [arXiv 2610.05608](https://arxiv.org/abs/2610.05608) (submitted 4 October 2026, updated 22 October), runs a pretrained video transformer stream and a separately trained audio stream side by side, connected by bidirectional cross-attention, so neither modality is a post-process on the other. That buys synchronised 44 kHz audio with lip sync. It does not buy length, and Full HD 1920×1080 comes from a separate super-resolution model.



Why Kandinsky runs a second stream instead of a louder first one

The naive way to put sound on generated video is to generate the video, then condition an audio model on it. That pipeline has an obvious failure mode: the audio model sees a finished pixel sequence and has to infer timing from it, so onsets drift and lip sync becomes a matter of luck and post-hoc alignment.

CrossDiT inverts the dependency. The audio stream is trained from scratch in its own right, then the two streams attend to each other during joint training, so timing information flows both ways. A plosive in the audio latent can influence the mouth shape in the video latent, and a footfall in the video can influence the transient in the audio. Structurally this resembles MMDiT in Stable Diffusion 3 and FLUX (separate parameter sets per modality, joined by attention), with a different purpose. MMDiT separates image latents from text tokens, which is a conditioning problem. CrossDiT separates video from audio, which is a synchronisation problem.

The analogy I would use for an engineer coming from distributed systems: MMDiT is a service reading from a config store, CrossDiT is two services holding a bidirectional stream and agreeing on a clock. The second one only works if both ends are first-class.

Four training phases and a flow-matching velocity field

The paper describes four phases. Audio-stream pretraining from scratch comes first, which matters because it means the audio stream is not a thin adapter bolted onto a video backbone. Then joint audio-video continuous pretraining on paired data, where the cross-attention links actually learn what to align. Then supervised fine-tuning. Then RL post-training plus distillation.

The sampling side is where the practical numbers are. Kandinsky 6.0 predicts a flow-matching velocity field rather than doing conventional discrete epsilon-prediction denoising. Flow matching defines a straighter path between noise and data, which is what makes aggressive step reduction viable at all. The distilled checkpoints use pi-Flow plus Sim-LADD and are advertised at roughly 10 NFE per sample: ten forward passes through



the transformer to produce a clip, instead of the dozens a classical diffusion schedule would want.

Ten NFE on a 29B dual-stream model is still ten passes through 29B parameters plus an audio stream plus whatever the super-resolution stage costs. I would treat the NFE figure as a comparative claim about sampler efficiency and not as a latency estimate, because it says nothing about batch size, precision, or the hardware underneath. A neural function evaluation is one call to the denoiser network during sampling, which makes it useful for comparing samplers and useless for predicting wall-clock time.

Video Lite at 3B is plausible on a workstation, Video Pro at 29B is not

The research data gives no VRAM figures for Kandinsky 6.0, so I will not invent any. What it does give is the shape of the decision: two sizes, 3B for Video Lite and 29B for Video Pro, both MIT licensed with Diffusers integration from day one. A 3B model is in the territory where a single workstation card is plausible. A 29B dual-stream model is not, and the roughly tenfold gap between the two sizes looks like a deliberate product split rather than a research artefact.

The neighbouring data point is more concrete. Lightricks' LTX-2.5, described on [the official model page](#) as a 22B-parameter asymmetric dual-stream diffusion transformer producing synchronised audio and video in a single pass, reports 80GB+ VRAM for the full model, with distilled and FP8 variants supporting 32GB and some configurations down to 12GB. Its headline speed number, 6.8 seconds for a 10-second 24fps clip, comes from [a RunPod benchmark](#) on two NVIDIA GB200 GPUs. That is a datacentre result, not something a consumer card reproduces.

My reading: if LTX-2.5 at 22B needs 80GB+ unquantised, a 29B Kandinsky Pro sits in the same bracket or worse, and the honest self-hosting path for most teams runs through distilled or FP8 checkpoints with the quality loss that implies. That is a different conversation from "the weights are MIT, so we can self-host", which is where a lot of procurement discussions stop. **The licence tells you what you are allowed to do, the GPU bill tells you what you will actually do**, a point I made at length in [a guide to self-hosted](#)



[ComfyUI pipelines.](#)

The five-second ceiling turns every product idea into a stitching problem

Kandinsky 6.0 outputs 5-second text-to-audio-video or image-to-audio-video clips. Anything that needs a 30-second sequence becomes a stitching problem, and stitching audio across clip boundaries without an audible seam is harder than stitching video. The architecture solves alignment inside a clip and says nothing about alignment between clips.

Three labs reached the same design inside one fortnight

The convergence is the part worth noticing. Kandinsky 6.0 on 6 October. LTX-2.5 at 22B with its asymmetric dual-stream design. Black Forest Labs making [FLUX 3 Image](#) generally available on 1 October 2026 as a jointly trained image/video/audio family built on flow matching and transformers, with the 5.6B open-weight FLUX 3 Image Klein following on 15 October. Three independent groups, inside the same fortnight, shipping flow matching plus separated modality streams. When the design space collapses that fast, it usually means the alternatives were tried and lost.

The regulatory side is less settled than the architecture. EU AI Act Article 50(2) requires machine-readable marking of synthetic audio, image and video “as far as technically feasible”. The provision is technology-neutral and names neither C2PA nor SynthID.

For a self-hosted generator that is your problem, not a vendor’s. An MIT-licensed checkpoint arrives with no marking obligation discharged on your behalf and no provenance metadata by default. You own the output and therefore the labelling. Marking generated audio is genuinely harder than marking generated images, which is the awkward corollary of an architecture that produces both in one pass: Article 50(2) covers audio explicitly, and the self-hosted path gives you nothing out of the box. I have written before about [how fragile C2PA content credentials are once files pass through platforms](#), and audio inherits those problems plus its own.



*My take: CrossDiT is the right architecture for the sync problem, and I expect the dual-stream pattern to be standard across open video models within two release cycles, because three labs reached it independently in October 2026. But I would not plan a 2027 product around self-hosting Video Pro. **The MIT licence on 29B weights is a research gift, not a deployment plan.** The realistic self-hosted option is the 3B Lite checkpoint or a distilled variant, and the realistic enterprise option is still an API where someone else owns the GPUs and the Article 50 marking. Unconfirmed, since the research gives no Kandinsky VRAM figures: I am extrapolating the hardware bracket from LTX-2.5's published 80GB+ at 22B. Treat that as inference.*

Three gaps the CrossDiT paper leaves open

The data does not describe what happens when the two streams disagree. Bidirectional cross-attention enforces alignment during training; it does not guarantee that a prompt demanding visually quiet footage with loud audio resolves sensibly rather than compromising both. That is an empirical question the published material does not answer.

The distillation claim and the quality claim are not stated together. Roughly 10 NFE is a sampler achievement, pi-Flow plus Sim-LADD is the method, and the research data contains no quality delta between the distilled and full checkpoints. For most teams the distilled checkpoint is the only one they will run, so that delta is the number that decides whether the model is useful.

The super-resolution stage is a separate model. The 1920×1080 output everyone will show in demos passes through a second network after the dual-stream transformer has done its work. That stage can introduce temporal artefacts the base model never produced, and it is a second set of weights to host, version and reason about.

None of this makes CrossDiT less interesting. It makes the gap between a fully inspectable architecture and a thing you ship wider than the release notes suggest. If you are weighing a self-hosted dual-stream video model against an API, that is a conversation I am happy to have, and [the fastest way to start it is just to write to me](#) with your current stack and your output volume.



CrossDiT, Explained: How Kandinsky 6.0 Generates Video and 44 kHz Audio in One Pass, and Where It Breaks



WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams turn AI from a chatbot into a reliable operational layer.

Book a meeting →