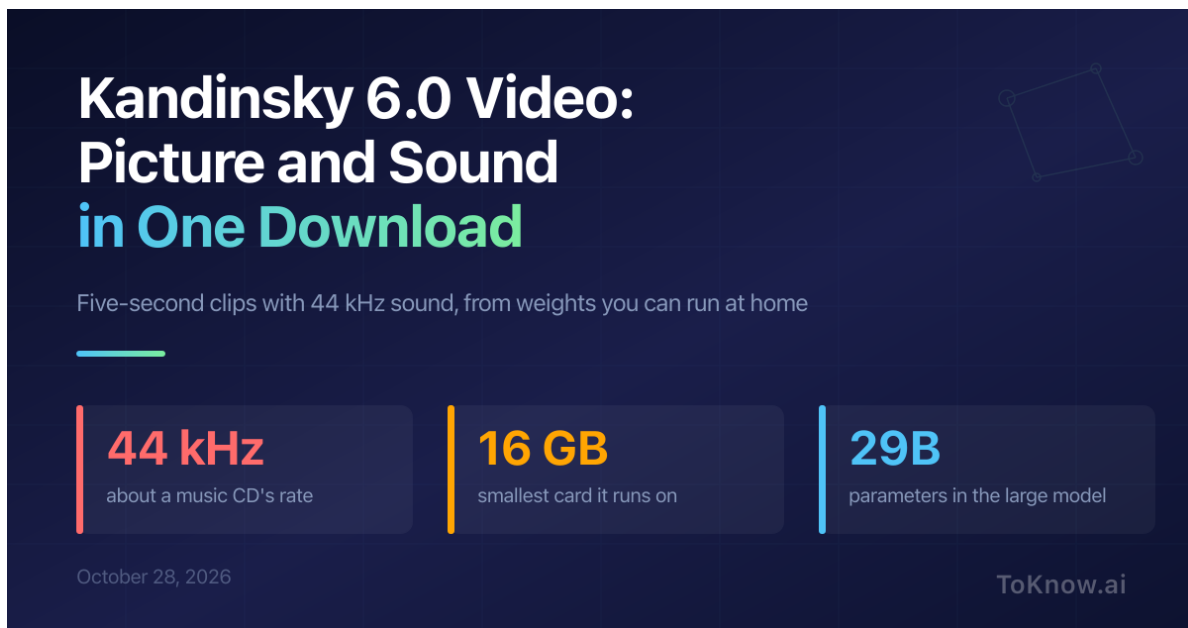


Kandinsky 6.0 Video: Picture and Sound in One Download

Kabui, Charles

2026-10-28

 Read at ToKnow.ai



The graphic is a dark blue rectangular banner. On the left, the title 'Kandinsky 6.0 Video: Picture and Sound in One Download' is written in white and light blue. Below the title, a line of text reads 'Five-second clips with 44 kHz sound, from weights you can run at home'. In the center, there are three dark blue boxes with vertical bars on their left sides. The first box has a red bar and contains the text '44 kHz' in red, with 'about a music CD's rate' below it. The second box has a yellow bar and contains '16 GB' in yellow, with 'smallest card it runs on' below it. The third box has a light blue bar and contains '29B' in light blue, with 'parameters in the large model' below it. On the right side of the banner, there is a faint, light blue geometric shape resembling a square with internal lines. At the bottom left, the date 'October 28, 2026' is written in small white text. At the bottom right, the 'ToKnow.ai' logo is displayed in white.

**Kandinsky 6.0 Video:
Picture and Sound
in One Download**

Five-second clips with 44 kHz sound, from weights you can run at home

44 kHz
about a music CD's rate

16 GB
smallest card it runs on

29B
parameters in the large model

October 28, 2026

ToKnow.ai

Kandinsky Lab released [Kandinsky 6.0 Video](#) on 6 October: a pair of downloadable models that make a five-second clip and its soundtrack in one pass. The small version has 3 billion parameters, the large one 29 billion. Both produce 44 kHz audio, about the rate of a music CD, with the lips timed to the speech. Picture and sound are trained together and exchange

signals, so the model learns the timing. A second model upscales the output to 1920 by 1080. The weights are free to download.

On [VABench](#), which scores generated sound and picture, the large model beat LTX 2.5, another open model, on lip-sync and audio-video alignment. Training that rewards clear speech cut the share of mistranscribed words by 47%. It also runs on a 16 GB graphics card, slowly: a Full HD clip from the full large model takes 26 minutes on an RTX 5080 and 59 on an RTX 5060 Ti, against 14 on an RTX 5090, and that setting needs about 60 GB of system memory.

In blind side-by-side tests, raters preferred it to LTX 2.5 on visuals and speech. Against paid systems the results are mixed: a split with Kling 2.6 and Veo 3.1 Fast, and MiniMax H3 and Seedance 2.0 ahead on visual quality. Clips stop at five seconds, against the 30 seconds [Wan 3.0](#) manages. It can also start from an audio file or from a first and last frame.

Read More: [daVinci-MagiHuman: One Model Generates Synchronized Video and Audio in 2 Seconds](#)

Sources:

- [Kandinsky 6.0 Video: Foundation Models for Synchronized Video and Audio Generation \(arXiv\)](#)
- [Full paper with the VABench and human-evaluation tables \(arXiv HTML\)](#)
- [kandinskylab/kandinsky-6: code, checkpoints and generation modes](#)
- [Kandinsky 6.0 Pro distilled on Hugging Face, 10-step sampling](#)

Disclaimer: For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation. **Read more:** [/terms-of-service](#)