daVinci-MagiHuman: One Model Generates Synchronized Video and Audio in 2 Seconds

Kabui, Charles

2026-03-29

Read at ToKnow.ai

SII-GAIR and Sand.ai released daVinci-MagiHuman, a 15B-parameter open-source model that jointly generates synchronized video and audio from a single text prompt. Most audio-video systems use complex multi-stream architectures with cross-attention to keep different outputs in sync. daVinci-MagiHuman uses a single-stream Transformer instead: text, video, and audio are all processed as one unified token sequence using only self-attention. A "sandwich" layout

gives the first and last four layers modality-specific projections while the middle 32 layers share parameters across all modalities. Combined with model distillation (down to 8 denoising steps), latent-space super-resolution, and a Turbo VAE decoder, the system generates a 5-second 256p video in 2 seconds on a single H100 GPU. In human evaluation across 2,000 pairwise comparisons, it achieved an 80.0% win rate against Ovi 1.1 and 60.9% against LTX 2.3, with the lowest word error rate (14.60%) for speech intelligibility among leading open models. It supports six languages: English, Mandarin, Cantonese, Japanese, Korean, German, and French.

That 2-second generation time matters. Near-real-time audio-video generation means content creators and developers can iterate on talking-head videos, multilingual voiceovers, or virtual assistant prototypes without waiting minutes per clip. The full stack is open-sourced under Apache 2.0, including the base model, distilled model, super-resolution model, and inference code, so anyone with an H100 (or even consumer GPUs with some adjustments) can run it.

The broader pattern here is striking: simpler architectures keep winning. Cross-attention, multi-stream coordination, and modality-specific pipelines are giving way to unified single-stream designs. ByteDance's Seedance took a similar multimodal approach, but daVinci-MagiHuman pushes architectural simplicity further while matching or exceeding quality.

Sources:

- daVinci-MagiHuman Paper (arXiv)
- daVinci-MagiHuman GitHub (1.1k stars)
- HuggingFace Model Page
- HuggingFace Paper Discussion (115 upvotes)
- HuggingFace Demo Space
