VoxCPM2: Tokenizer-Free TTS Generates 48kHz Speech in 30 Languages Without an Audio Codec

Kabui, Charles

2026-06-06

Tsinghua University's OpenBMB lab released VoxCPM2, a 2-billion-parameter text-to-speech model that removes the audio tokenizer from the synthesis pipeline entirely. Most modern TTS systems convert text to discrete audio tokens using a codec like EnCodec or DAC, then

decode those tokens back into a waveform. That quantization step loses information, and the codec becomes a quality ceiling. VoxCPM2 replaces it with a diffusion autoregressive architecture that operates in continuous latent space, generating 48kHz audio directly. Built on a MiniCPM-4 backbone and trained on over 2 million hours of multilingual speech, it supports 30 languages including 9 Chinese dialects. You can clone any voice from a short reference clip, or skip reference audio and describe the voice you want in plain text: "(young woman, warm and gentle)" before the words to speak.

VoxCPM2 runs at a real-time factor of about 0.3 on an RTX 4090, dropping to roughly 0.13 with the Nano-vLLM serving engine, and needs around 8 GB of VRAM. Fine-tuning to a specific voice takes as little as 5 to 10 minutes of audio. Everything ships under Apache 2.0, so commercial use requires no separate license. A filmmaker who needs narration in Thai, a game studio building NPCs, or a localization team dubbing across dozens of languages can run one model for all of it.

VoxCPM2 sits at one end of a growing architectural split in TTS. Systems like Fish Audio S2 use discrete token pipelines and achieve top-tier quality. VoxCPM2 argues the tokenizer itself is the bottleneck and removes it. Both are now open-source and commercially usable, which shifts the competition from access to architecture.

Sources:

- VoxCPM2 GitHub Repository (OpenBMB)
- VoxCPM Technical Report (arXiv)
- VoxCPM2 Model Weights (Hugging Face)
- VoxCPM2 Demo Page and Audio Samples
