Meituan LongCat-Next: One Model That Sees, Draws, and Talks Using a Single Token System

Kabui, Charles

2026-04-04

Meituan's LongCat team released LongCat-Next, a 74B-parameter open-source model that processes text, images, and audio through a single system. Rather than bolting separate vision and audio modules onto a language model, LongCat-Next converts all modalities into

discrete tokens in a shared space, a framework the team calls DiNA (Discrete Native Autoregression). The core technical piece is dNaViT, a visual tokenizer that converts images at any resolution into hierarchical "visual words" using semantic encoders paired with residual vector quantization. This preserves both high-level meaning and fine pixel-level detail. One model handles image understanding, image generation, speech comprehension, voice conversation, and voice cloning under a single autoregressive objective, matching specialized understanding models while maintaining strong generation quality at a 28x compression ratio.

Previous discrete-token multimodal models consistently sacrificed understanding quality for generation ability, or the reverse. LongCat-Next reconciles that trade-off. Its Mixture-of-Experts backbone keeps only 3B parameters active at inference despite the 74B total, so it runs more efficiently than the raw number suggests. The full model, tokenizers, and inference code ship under an MIT license with over 5,600 downloads on HuggingFace already. Anyone building multimodal applications can fine-tune or deploy it without restrictions.

Multimodal AI is shifting from "language model with vision bolted on" to natively multimodal from the ground up. LongCat-Next, alongside recent releases like Qwen3.5, makes a strong case that treating vision and audio as first-class token types, not adapter outputs, produces better results across the board.

Sources:

- LongCat-Next arXiv paper
- LongCat-Next GitHub repository
- LongCat-Next on HuggingFace
- LongCat-Next project page
