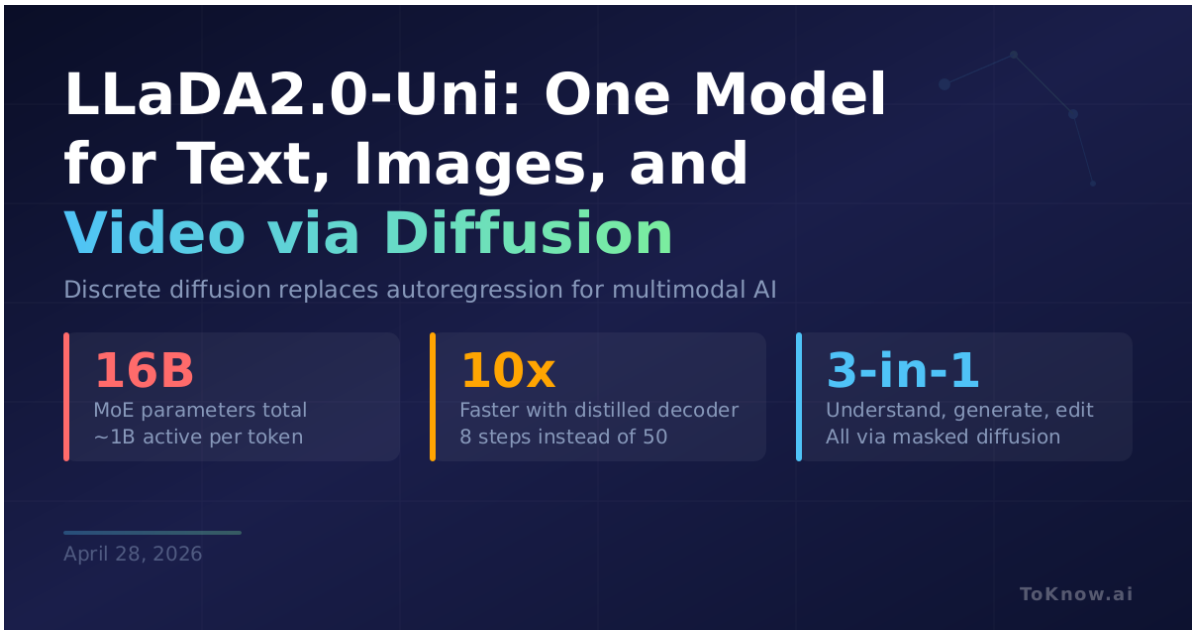
LLaDA2.0-Uni: One Diffusion Model for Understanding and Generating Text, Images, and Video

Kabui, Charles

2026-04-28



Inclusion AI released LLaDA2.0-Uni, a multimodal model that uses discrete diffusion instead of autoregressive generation to both understand and create text, images, and video. The architecture combines three parts: a visual tokenizer (SigLIP-VQ) that converts images into

discrete tokens identical to text tokens, a 16B Mixture-of-Experts backbone that processes all modalities through masked diffusion while activating only ~1B parameters per token, and a diffusion decoder for image reconstruction. Because text and images share one token space, the same model handles image captioning, text-to-image generation, and image editing through different masking patterns. A distilled decoder cuts image generation from 50 steps to 8, roughly 10× faster. The full pipeline needs about 47 GB of VRAM; understanding alone fits in 35 GB.

Most multimodal models bolt image generation onto a language backbone as a separate system. LLaDA2.0-Uni treats both as the same operation: denoise masked tokens. One model, one training objective, one inference stack handles everything from answering questions about a photo to generating new images from text. The Apache 2.0 license and full weight release make it accessible for any research team. Understanding-only mode fits on a single 40 GB GPU, keeping experimentation practical outside massive compute clusters.

Autoregressive generation has dominated LLMs since GPT-2. Continuous diffusion dominates image synthesis. LLaDA2.0-Uni, alongside Meituan's LongCat-Next, represents a growing bet that discrete diffusion can unify both. If the approach scales, future foundation models may drop autoregression entirely, generating all tokens in parallel rather than one at a time.

Sources:

- LLaDA2.0-Uni Technical Report (arXiv)
- LLaDA2.0-Uni GitHub Repository
- LLaDA2.0-Uni Model Card (HuggingFace)
- LLaDA2.0-Uni Paper Discussion (HuggingFace)
