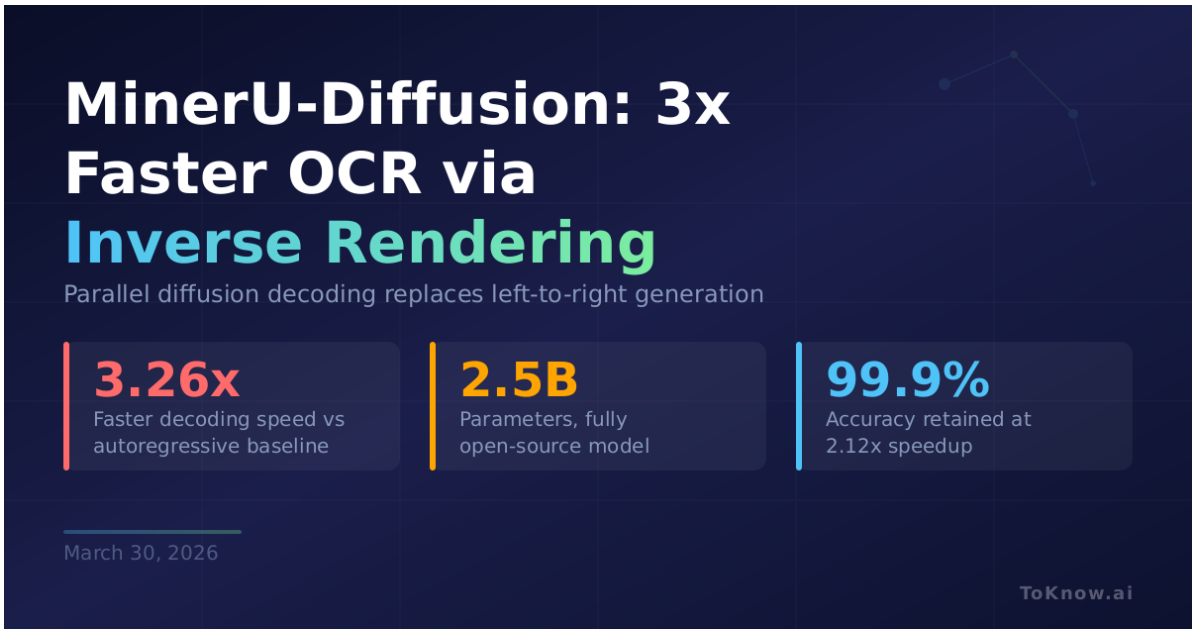
MinerU-Diffusion: Document OCR Rethought as Inverse Rendering, 3x Faster

Kabui, Charles

2026-03-30

Read at ToKnow.ai



OpenDataLab released MinerU-Diffusion, a 2.5B-parameter document OCR model that replaces the standard left-to-right text generation used by most systems with parallel diffusion decoding. The core insight: recovering structured content from a document image is an inverse rendering problem, not a sequential prediction task. Documents are 2D spatial layouts, so generating their content one token at a time is an artificial constraint. The model splits

its output into blocks where tokens generate simultaneously, while a sequential scaffold across blocks preserves global coherence. It achieves up to 3.26x faster decoding compared to its autoregressive predecessor, with practical sweet spots at 2.12x speedup retaining 99.9% accuracy and 3.01x at 98.8%. A new Semantic Shuffle benchmark confirms the model reads the image rather than guessing from language patterns.

Enterprise document pipelines process millions of pages. A 3x speed improvement at near-perfect accuracy directly cuts compute costs and turnaround for extracting tables, formulas, and structured layouts from scanned documents. The reduced dependency on language patterns also matters for multilingual documents and non-Latin scripts, where left-to-right models tend to autocomplete from what they expect linguistically instead of reading what's actually on the page.

Autoregressive decoding has dominated vision-language models because it works and the tooling is mature. MinerU-Diffusion suggests that for spatial tasks like document parsing, the sequential assumption is the bottleneck, not model size. If this holds, the same inverse rendering logic could reshape how models handle diagrams, sheet music, and architectural drawings.

Sources:

- MinerU-Diffusion Paper (arXiv)
- MinerU-Diffusion GitHub Repository
- MinerU-Diffusion Model on HuggingFace
- MinerU-Diffusion Demo
