NVIDIA LocateAnything: Spotting Objects 10x Faster By Predicting Whole Bounding Boxes At Once

Kabui, Charles

2026-06-01

NVIDIA Research released LocateAnything, a 3-billion-parameter vision-language model that predicts each bounding box in a single step instead of emitting its four coordinates as separate tokens. Standard vision-language models serialize a 2D box into a number sequence

and decode it sequentially, a slow process that ignores the geometric coupling of the corners. LocateAnything's Parallel Box Decoding emits the full coordinate set atomically. The model pairs a Moon-ViT vision encoder with a Qwen2.5 language decoder and was trained on 138 million queries and 785 million boxes. On a single H100 it outputs 12.7 boxes per second, roughly **10x faster** than Qwen3-VL (1.1) and 2.5x faster than Rex-Omni (5.0), and lifts F1 on the LVIS detection benchmark by +3.8% over Rex-Omni, with high-overlap accuracy (IoU=0.95) widening from 20.7 to 31.1.

That combination unlocks live use of vision-language models where they previously had to run offline. A warehouse robot scanning a shelf or a self-checkout reading 50 items can now run a VLM each frame. GUI agents driving Windows or web apps reach state-of-the-art element grounding (60.3 mean F1 on ScreenSpot-Pro), and document parsing reaches 76.8 mean F1 on DocLayNet, cutting invoice and contract extraction costs at scale. A Hybrid mode falls back to sequential decoding only when a parallel output looks malformed, so robustness stays intact.

LocateAnything argues the bottleneck for visual grounding is the output format itself, not model size or training data. Treat each spatial unit as atomic instead of a 1D token stream and you gain speed and accuracy together. The same shift is happening in parallel diffusion decoding for document OCR: when the task is spatial, sequential decoding is the real constraint.

Sources:

- LocateAnything project page (NVIDIA Research)
- LocateAnything paper (arXiv:2605.27365)
- LocateAnything technical PDF
- LocateAnything-3B model on HuggingFace
- LocateAnything code on GitHub (NVlabs/Eagle)
