

Now Available: SDK

vLLM Official Plugin, Compiler Optimization, and Model Zoo Expansion

Highlights

vLLM RBLN refactored and open-sourced as an official plugin



- Seamless upstream compatibility (vLLM v0.9.1)
- Support for various language models including decoder-only and encoder-decoder architectures, as well as up-to-date vision-language multimodal models

Learn more in the [vLLM RBLN documentation](#) or explore the [GitHub repository](#) for full details.

Compiler was enhanced to improve efficiency and optimize memory usage:

- Improved handling of diverse input shapes in Scaled Dot-Product Attention (SDPA) kernel to enable more efficient and flexible attention
 - Reduced device DRAM footprint by optimizing memory handling on graph breaks
 - Faster model builds with lower host DRAM consumption for large-scale models
-

New HuggingFace 🤖 models added to the Model Zoo:

- | | |
|--|--|
| ▪ Cosmos-1.0-Diffusion-7B/14B-Text2World  | ▪ A.X-4.0-Light  |
| ▪ Cosmos-1.0-Diffusion-7B/14B-Video2World  | ▪ Midm-2.0-Mini/Base  |
| ▪ Gemma3-4B/12B/27B  | |

Explore all supported models in the [RBLN Model Zoo](#).

Documentation restructured by domain for improved clarity:

- Software (Basic Tools) — RBLN Compiler, RBLN Optimum, RBLN Profiler
- Model Serving — vLLM, Triton Inference Server, TorchServe
- Cloud-Native Support — Kubernetes Support (Device Plugin, NPU Feature Discovery, Metrics Exporter), System Management

For full details on this month's updates, check out [the official release notes](#). We're releasing the RBLN SDK monthly, and we'll continue to deliver faster performance, broader model support, and an improved developer experience — stay tuned!