精选快讯　软件快讯　2026年11月19日 13:11

辉羲智能开源具身智能推理引擎 RhinoForge，支持自主适配与跨模型复用

要点

- 辉羲智能正式开源具身智能推理引擎 RhinoForge。
- RhinoForge 采用接口层、模型适配层、运行时层和算子层的四层解耦设计，将自研 RPU 计算内核直接接入为 PyTorch 可调用的计算设备。
- 在模型适配方面，RhinoForge 已支持 Pi0.5、WALL-OSS、G0.5、GR00T、RhinoVLA、LingBot VLA 2.0 等 VLA 模型，以及 Qwen3-VL、DINOv3、Qwen3-ASR/TTS 等多模态、视觉与语音模型和 Qwen3、Qwen3.5 等大语言模型，并面向 OpenWAM、Cosmos 等世界模型提供适配支持。
- 底层硬件方面，Embodied R1 系列芯片采用通用 SIMT 可编程计算架构，具备图灵完备的指令集。

辉羲智能正式开源具身智能推理引擎 RhinoForge。该框架面向自研 RPU 计算内核，旨在提供 PyTorch 推理与模型部署能力，目前支持 Huixi Embodied R1 MAX 和 Huixi Embodied R1 PRO 两款具身智能大算力芯片。

RhinoForge 采用接口层、模型适配层、运行时层和算子层的四层解耦设计，将自研 RPU 计算内核直接接入为 PyTorch 可调用的计算设备。开发者可使用高级语言 rpuc 编写新算子，经 HXCC 编译后接回 PyTorch，从而自主适配模型并扩展计算算子，将底层优化能力沉淀为可跨模型复用的工程资产。

在模型适配方面，RhinoForge 已支持 Pi0.5、WALL-OSS、G0.5、GR00T、RhinoVLA、LingBot VLA 2.0 等 VLA 模型，以及 Qwen3-VL、DINOv3、Qwen3-ASR/TTS 等多模态、视觉与语音模型和 Qwen3、Qwen3.5 等大语言模型，并面向 OpenWAM、Cosmos 等世界模型提供适配支持。该引擎支持 INT8、FP8、FP16/BF16、INT4、NVFP4 等量化方案，允许按模块敏感度组合使用，并支持执行图缓存与重以优化推理性能。

底层硬件方面，Embodied R1 系列芯片采用通用 SIMT 可编程计算架构，具备图灵完备的指令集。在辉羲智能研发团队基于 OpenPI 格式的 LIBERO 测试中，视觉 - 语言 - 动作模型 Pi0.5 在 Embodied R1 MAX 芯片上的端到端推理频率达到 12.5 Hz。

报道来源：jushenhome

来源网站：https://www.inews.today/