

PR #28481 完整报告

sgl-project/sglang

chore: bump sgl-deep-gemm build-time apache-tvm-ffi 0.1.9 -> 0.1.11

合并时间: 2026-06-28 15:26

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28481>

执行摘要

- 一句话: 升级 DeepGEMM 构建时 tvml-ffi 版本
- 推荐动作: 值得关注其作为分步升级流程的一部分, 体现了依赖版本对齐的工程实践。但本身不包含设计决策或深度技术内容。

功能与动机

与 sgl-project/DeepGEMM#47 配合, 将 sgl-deep-gemm wheel 的 runtime tvml-ffi 升级到 0.1.11。构建时和运行时的 tvml-ffi 必须一致, 否则已发布的 wheel 链接一个 ABI 却解析另一个 ABI。选择 0.1.11 而非最新 0.1.12 是为了规避 tile-ai/tilelang#2367 的 duplicate `__ffi_repr__` import-abort 回归问题。

实现拆解

变更仅涉及一个文件 `docker/sgl-deep-gemm.Dockerfile`, 将 `TVM_FFI_VER` 从 0.1.9 改为 0.1.11。该变量用于 `release-whl-deepgemm.yml` 工作流, 在编译 sgl-deep-gemm wheel 的 `_C.so` 时安装构建时 `apache-tvm-ffi`。此变更不影响当前安装, 只影响下一次 wheel 构建。

关键文件:

- `docker/sgl-deep-gemm.Dockerfile` (模块 构建配置; 类别 infra; 类型 infrastructure; 符号 `TVM_FFI_VER`): 唯一变更文件, 修改构建时 tvml-ffi 版本号

关键符号: 未识别

评论区精华

PR 无 review 评论。唯一的 review 来自 Fridge003, 直接批准 (APPROVED)。

- 暂无高价值评论线程

风险与影响

- 风险: 极低风险: 仅修改一个 Dockerfile 中的构建参数, 不影响任何运行时逻辑。由于 0.1.11 已在 Hopper 和 Blackwell 平台上经过端到端验证 (DeepGEMM 测试套件和 SGLang 服务均通过), 并且下轮构建才会生效, 即使出问题也不会影响现有部署。
- 影响: 影响范围有限, 仅影响未来 sgl-deep-gemm wheel 的构建产物。对现有用户无直接感知, 但确保了新 wheel 的 ABI 兼容性, 为后续全面升级铺路。

- 风险标记: 暂无

关联脉络

- PR #47 chore: bump apache-tvm-ffi 0.1.9 -> 0.1.11: sgl-project/DeepGEMM 中的对应 PR, 升级 wheel 运行时依赖至 0.1.11