

PR #47405 完整报告

vllm-project/vllm

[XPU][CI]Mv huggingface cache to larger disk in Intel GPU CI

合并时间: 2026-07-03 11:56

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/47405>

执行摘要

- 一句话: 转移 Intel GPU CI 的 HuggingFace 缓存到更大磁盘
- 推荐动作: 该 PR 为基础设施微调, 变更简单明确, 风险低, 适合快速合并。无需深入 review。

功能与动机

Intel GPU CI 的 HuggingFace 缓存默认存储在用户 home 目录下, 磁盘空间较小, 容易因模型下载缓存耗尽空间导致测试失败。将其挂载到 `/data` 路径下的更大磁盘可缓解该问题。

实现拆解

1. 修改 Docker 挂载卷路径 (`.buildkite/scripts/hardware_ci/run-intel-test.sh`): 将容器内 HuggingFace 缓存目录 `/root/.cache/huggingface` 挂载到宿主机的 `/data/huggingface` 而非原来的 `${HOME}/.cache/huggingface`。
2. 更新 CI 全量触发配置 (`.buildkite/ci_config_intel.yaml`): 在 `run_all_patterns` 列表中添加 `.buildkite/scripts/hardware_ci/run-intel-test.sh`, 使得对该脚本的修改能够触发 Intel CI 的全量构建和测试, 而不会漏过。

关键文件:

- `.buildkite/scripts/hardware_ci/run-intel-test.sh` (模块 部署脚本; 类别 `infra`; 类型 `infrastructure`): 修改 Docker 挂载卷路径, 将 HuggingFace 缓存从 `home` 目录迁移到更大磁盘的 `/data/huggingface` 目录。
- `.buildkite/ci_config_intel.yaml` (模块 部署脚本; 类别 `config`; 类型 `configuration`): 在 `run_all_patterns` 中添加 `run-intel-test.sh`, 确保该脚本变更能触发全量 CI。

关键符号: 未识别

评论区精华

无人工 review 讨论。仅 bot 留言表示 fork 仓库无法自动 review, 随后仓库维护者 jikunshang 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：
 - 路径不存在风险：若 `/data/huggingface` 目录在宿主机上不存在，Docker 挂载会失败。但该目录在 Intel GPU CI 环境中预期已存在。
 - 缓存初始化：首次迁移后，新路径下可能没有预下载模型缓存，可能导致首个 CI 任务需要重新下载模型，增加时间开销。
 - 配置完整性：`run_intel_test.sh` 已加入 `run_all_patterns`，配置无遗漏。
- 影响：
 - 仅影响 Intel GPU CI 环境，不影响其他 CI 或生产代码。
 - CI 稳定性提升：避免因磁盘不足导致的假性失败。
 - 首次运行可能变慢：新缓存路径为空时，首次 CI 需要重新下载模型。
 - 风险标记：依赖外部磁盘路径，首次运行需重新下载缓存

关联脉络

- PR #46456 [CI] intel CI: add quantization and awq case for xpu: 同为 Intel GPU CI 基础设施变更，涉及同一脚本和配置目录。
- PR #47510 [XPU][CI]Fix dependency typo in Intel GPU CI: 同为 Intel GPU CI 修复，涉及同一 CI 配置目录。