

PR #56704 完整报告

vllm-project/vllm

[XPU][CI] update test requirements

合并时间: 2026-09-14 10:01

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/56704>

执行摘要

- 一句话: XPU 测试依赖对齐 cuda.in, 修复 gpt-oss-20b CI 失败
- 推荐动作: 不值得精读, 可作为 vLLM 平台化测试依赖管理的快速参考。值得关注的要点: requirements/test 下按平台拆分 .in 文件的做法、gpt-oss 这类模型评估包通过版本下限加环境标记 (python_version、platform_machine) 做条件安装的写法, 以及 .in 与 .txt 的分工 (源声明与编译产物)。后续接手 XPU CI 维护时, 应保持 xpu.in 与 cuda.in 的同步习惯。

功能与动机

PR body 明确指出目的: align xpu test requirements with cuda.in resolve this ci error.

在 Buildkite Intel CI (build 11197) 上有 3 个测试失败:

entrypoints/openai/responses/test_harmony.py::test_code_interpreter[openai/gpt-oss-20b]、test_mcp_tool_multi_turn[openai/gpt-oss-20b], 以及

entrypoints/openai/responses/test_mcp_tools.py::TestMCPEnabled::test_mcp_tool_env_flag_enabled[openai/gpt-oss-20b]。三个失败用例的共同点是参数化 openai/gpt-oss-20b 模型, 说明 xpu.in 中 gpt-oss[eval]==0.0.5 的旧版本与 CUDA 侧测试依赖矩阵发生漂移, 导致工具调用与 MCP 相关测试在 XPU 环境无法正确执行。

实现拆解

变更入口是 requirements/test/xpu.in, xpu.txt 由其经 pip-compile 重新编译生成。具体步骤:

1. gpt-oss 升级并条件化: 将 gpt-oss[eval]==0.0.5 改为 gpt-oss >= 0.0.7; python_version > '3.11'。去掉 eval extra 并抬高版本下限, 同时限定 Python 版本, 避免旧版 Python 环境安装无关依赖; 这是修复 gpt-oss-20b 相关测试失败的关键动作。
2. 版本下限补齐: schemathesis>=4.0.0 (openai schema 测试需要)、timm>=1.0.17 (internvl 与 gemma3n-mm 测试需要)、mistral-common[image,audio]>=1.11.6 (voxtral 测试需要), 每项均附带注释说明用途, 防止后续被误降级。
3. 平台条件约束: decord==0.6.0 追加 platform_machine == "x86_64" 条件, 避免非 x86_64 平台因缺少 wheel 安装失败。
4. 重新生成锁定文件: xpu.txt 将 gpt-oss 锁定到 0.0.9, 并带入 jupyter-client、jupyter-core、tornado、traitlets 等传递依赖, 同时清理旧版 gpt-oss 的残留引用注释。

配套说明：无新增测试用例、无部署变更；两个 commit 中第二个为 merge main 的常规同步。

关键文件：

- requirements/test/xpu.in（模块 测试依赖；类别 test；类型 configuration）：本 PR 的变更入口：将 gpt-oss 升级为 $\geq 0.0.7$ （限定 Python > 3.11 ），为 schemathesis、timm、mistral-common 补充版本下限，给 decord 加 x86_64 平台条件，直接解决 3 个 gpt-oss-20b 相关测试失败，并与 cuda.in 对齐。
- requirements/test/xpu.txt（模块 测试依赖；类别 docs；类型 documentation）：pip-compile 生成的锁定文件，将 gpt-oss 锁定到 0.0.9 并带入 jupyter-client、jupyter-core、tornado、traitlets 等传递依赖，保证 CI 依赖可复现，也是核实 xpu.in 变更能成功解析的依据。

关键符号：未识别

评论区精华

本次 review 过程非常简短，没有实质性技术争论：

- claude[bot] 提示该 PR 来自 fork，自动 review 被禁用，维护者可评论 @claude review 触发一次性审查。
- 作者 @zhenwei-intel 请求 @jikunshang PTAL 后，维护者回复 /ci run 触发 Buildkite CI #88698，github-actions[bot] 确认在 commit 3831fa72759b 上运行。
- jikunshang 最终 APPROVED，原话为 thanks for fixing。
- fork PR 自动 review 禁用 (other)：维护者 jikunshang 手动审批通过 (thanks for fixing.)，未启用自动 review。
- CI 验证流程 (testing)：CI 在 head commit 上运行通过后 PR 被合并。

风险与影响

- 风险：
 1. 平台条件依赖：decord 限定仅 x86_64 安装后，非 x86_64 的 CI runner 将缺少该包；若相关视频解码测试没有 skip 防护，会直接出现 ModuleNotFoundError，需确认测试侧已有对应跳过逻辑。
 2. Python 版本条件：gpt-oss 仅在 python_version $> '3.11'$ 时安装，若 XPU CI 环境回退到 Python 3.10 及以下，gpt-oss-20b 相关测试仍会失败；本次 CI 通过以 3.11+ 环境为前提。
 3. 依赖升级传递风险：gpt-oss 0.0.7+ 引入 jupyter-client、jupyter-core、tornado、traitlets 等新传递依赖，未来与其它平台锁文件解析时可能产生版本冲突；本次 xpu.txt 已成功解析，但需与 cuda.txt 保持同步更新节奏。
 4. 运行时风险：无。变更不触碰 vllm 产品代码路径。
- 影响：
 - 用户：无直接影响，纯 CI 测试依赖基础设施调整。

- 系统：修复 Intel Buildkite 上 3 个 gpt-oss-20b 相关测试失败，XPU 测试矩阵与 CUDA 侧重新对齐。
- 团队：Intel CI 维护者后续需沿用统一模式：在 xpu.in 声明依赖并注释原因，再重新编译 xpu.txt；其他平台开发者也可参考本 PR 的分平台依赖写法（platform_machine、python_version 条件）。
- 风险标记：平台条件依赖，Python 版本条件限制，依赖升级传递风险，测试依赖漂移

关联脉络

- PR #50097 [XPU] fix_test_worker_memory_snapshot_for_xpu: 同属 XPU CI 测试基础设施维护线，说明 Intel 团队在持续修复 XPU 平台测试兼容性；本 PR 进一步对齐 xpu 与 cuda 的测试依赖矩阵。