

PR #55171 完整报告

vllm-project/vllm

[XPU][CI] Remove pip install dependency in test yaml files

合并时间: 2026-09-12 20:03

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/55171>

执行摘要

- 一句话: XPU CI 依赖统一收敛至 xpu.in, 8 处 pip install 下线
- 推荐动作:
 1. 是否值得精读: 对 Intel/XPU CI 维护者、以及负责 requirements 与镜像构建的同学值得精读; 对做模型、内核、调度的工程师可快速略过 (零运行时源码改动)。
 2. 值得关注的设计决策: 把“步骤级 pip install”整体上移到镜像 requirements, 是一次典型的 CI 依赖治理; 阅读时应重点对照删除命令与 xpu.in 新增条目的映射表, 确认没有遗漏 (例如 pip install open-clip-torch --no-deps 改为带依赖解析的 open_clip_torch==2.32.0 属于语义变化, 不只是形式迁移)。
 3. 建议的后续动作: 跟进补上 requirements/test/xpu.in、requirements/test/xpu.txt 到 .buildkite/intel_jobs/*.yaml 的 source_file_dependencies 及 .buildkite/ci_config_intel.yaml; 在 PR 描述中补测试证据 (至少说明依赖已随镜像生效); 对 gpt-oss 0.0.9 → 0.0.5、numpy 2.3.5 → 2.2.6 的锁文件回退给出验证结论或原因说明。

功能与动机

PR body 只有一句核心表述: “move all dependency into requirements/test/xpu.in file”。其动机是: XPU 的 5 个 intel job yaml 中长期存在大量步骤级 pip install (tblib、av、matplotlib、ftfy、open-clip-torch、tensorizer、runai-model-streamer、lm_eval、modelscope、opentelemetry 四件套、gpt-oss[eval]), 这些命令在每个测试步骤启动时都要联网重复解析和安装, 既拖慢 CI (每个 step 多出几十秒到数分钟), 又让依赖版本随上游漂移、难以复现。把这些依赖收进 XPU 测试镜像的 requirements 清单, 可以做到“镜像一次装好、步骤直接跑测试”, 与仓库中 requirements/.in +.txt 的既有依赖管理模式保持一致。

实现拆解

1. 依赖清单归拢 (requirements/test/xpu.in, +19/-3): 新增 tblib==3.1.0、runai-model-streamer[s3,gcs,azure]>=0.15.7、tensorizer==2.10.1、av、gpt-oss[eval]==0.0.5、ftfy、matplotlib、open_clip_torch==2.32.0、opentelemetry-api/exporter-otlp/sdk/semantic-conventions-ai 四个带版本下限的条目, 以及 cohere_melody>=0.11.1、cohere>=7.0.0 两个 parser/API 测试依赖; 同时把原先裸写的 tblib 固定为 tblib==3.1.0、裸 gpt-oss 改为 gpt-oss[eval]==0.0.5, 并补上文件末尾换行。

2. 锁文件重算 (requirements/test/xpu.txt, +172/-26) : 由 xpu.in 重新解析生成。新增链条包括 runai-model-streamer 的三个存储后端 (boto3/botocore/s3transfer、google-cloud-storage/google-auth/google-api-core、azure-identity/azure-storage-blob/msal)、tensorizer 的 hiredis/libnacl、matplotlib 的 contourpy/cycler/fonttools/kiwisolver/pyparsing、open-clip-torch 引入的 ftfy; gpt-oss 由 0.0.9 回退到 0.0.5 并连带移除 jupyter-client/jupyter-core, numpy 由 2.3.5 降到 2.2.6, av==18.1.0、cohere==7.1.1 为新增固定版本。
3. 流水线瘦身 (5 个 .buildkite/intel_jobs/*.yaml, 净 -18 行) : 删除的安装命令与其迁移目标一一对应: test-intel.yaml 去掉 pip install tblib==3.1.0 与 pip install av; misc_intel.yaml 去掉 lm_eval[api]>=0.4.12、modelscope<1.38、四条 opentelemetry 包与 av; models_multimodal_intel.yaml 去掉三处 av/matplotlib/ftfy 与 pip install open-clip-torch --no-deps; model_executor_intel.yaml 去掉 tensorizer==2.10.1 与 runai-model-streamer[s3,gcs,azure]>=0.15.7; lm_eval_intel.yaml 去掉 gpt-oss[eval]==0.0.5。改造后这些命令统一以 cd tests && pytest ... 开头, export 类环境变量 (如 VLLM_WORKER_MULTIPROC_METHOD=spawn) 保留在命令串内。
4. 配套与缺口: gpt-oss[eval] 的 pin 从 yaml 迁到 xpu.in 后, 锁文件里实际降级为 0.0.5, 与 lm_eval 步骤原有 pin 对齐, 但会影响其他依赖 gpt-oss 的用例; 更关键的是 CodeRabbit 指出 .buildkite/intel_jobs/test-intel.yaml 等 job 的 source_file_dependencies 以及 .buildkite/ci_config_intel.yaml 均未加入 requirements/test/xpu.in 与 requirements/test/xpu.txt, 导致后续只改依赖文件不会触发对应 XPU 测试。该 PR 未补测试代码, 验证完全依赖两次 /ci run (Buildkite #88142、#88489)。

关键文件:

- requirements/test/xpu.in (模块 测试依赖; 类别 test; 类型 dependency-manifest; 符号 tblib==3.1.0, runai-model-streamer[s3,gcs,azure], tensorizer==2.10.1, av) : 本次变更的单一依赖真源: 所有从 5 个 Buildkite job 里删除的 pip install 都在此固化, 并新增 pin 与版本下限, 是理解整份 PR 的入口文件。
- requirements/test/xpu.txt (模块 依赖锁文件; 类别 config; 类型 dependency-lock; 符号 gpt-oss==0.0.5, numpy==2.2.6, av==18.1.0, open-clip-torch==2.32.0) : 由 xpu.in 重算的锁文件 (+172/-26), 把版本漂移风险与新增传递依赖 (S3/GCS/Azure 后端、tensorizer、matplotlib) 全部显性化, 也是“依赖文件变更应触发 XPU 测试”的对象之一。
- .buildkite/intel_jobs/test-intel.yaml (模块 流水线配置; 类别 config; 类型 configuration; 符号 xpu-example-test, xpu-server-test) : 删除 tblib 安装的落地文件, 同时也是 CodeRabbit 提出 source_file_dependencies 覆盖缺口的具体位置 (第 37 行附近), 代码评审的主要争议点在此。
- .buildkite/intel_jobs/misc_intel.yaml (模块 流水线配置; 类别 config; 类型 configuration; 符号 async-engine-inputs-utils-worker, metrics-tracing-2-gpus, regression) : 删除条目最多 (lm_eval、modelscope、四条 opentelemetry 包、av), 覆盖 Basic Models / Regression / Metrics-Tracing / Async Engine 四个 job, 是验证依赖迁移完整性的重点文件。

- .buildkite/intel_jobs/models_multimodal_intel.yaml (模块 流水线配置; 类别 config; 类型 configuration; 符号 multi-modal-processor, multi-modal-models-standard-1-qwen2) : 删除三处安装命令 (av、matplotlib/ftfy、open-clip-torch --no-deps), 后者从 --no-deps 改为带依赖解析, 是本次最容易被忽略的语义变化点。
- .buildkite/intel_jobs/model_executor_intel.yaml (模块 流水线配置; 类别 config; 类型 configuration; 符号 model-executor-intel) : 删除 tensorizer 与 runai-model-streamer 的安装, 这两个包正是 xpu.txt 中新增依赖链 (boto3/azure/google-cloud-storage/hiredis) 的源头。
- .buildkite/intel_jobs/lm_eval_intel.yaml (模块 流水线配置; 类别 config; 类型 configuration; 符号 xpu-gpqa-eval-gpt-oss) : 删除 gpt-oss[eval]==0.0.5 安装, 该 pin 被迁入 xpu.in 并导致锁文件中 gpt-oss 版本变化, 是版本回退风险的来源之一。

关键符号: xpu-example-test (Buildkite step key), xpu-server-test, model-executor-intel, xpu-gpqa-eval-gpt-oss, multi-modal-processor, metrics-tracing-2-gpus

评论区精华

1. CodeRabbit 的唯一 actionable 评论 (未解决) : 在 .buildkite/intel_jobs/test-intel.yaml 第 37 行提出“Update Intel CI trigger coverage by adding requirements/test/xpu.in and requirements/test/xpu.txt to the relevant source_file_dependencies and . buildkite/ci_config_intel.yaml configuration, ensuring changes to either XPU test requirements file schedule the affected XPU tests”, 并附了 rg 'source_file_dependencies!image-build-xpulrequirements/test/xpu\.(in|txt)' 的静态分析脚本输出作为证据。PR 只提交了 remove pip install 与 merge main 两个 commit, 该评论未见回复、也未在后续 commit 中修复。
 2. 人工评审只有一次 APPROVED: reviewer bigPYJ1151 直接 approve, 没有留下文字意见, 说明 Intel CI 侧维护者认可这次收敛方向。
 3. claude[bot] 因 fork 来源禁用自动评审: “This pull request is from a fork — automated review is disabled.”, 因此本 PR 实质只经过 CodeRabbit 一轮自动审查, 评审深度有限。
 4. 协作流程: 作者在评论中发起两次 /ci run (Buildkite #88142、#88489), 期间出现合并冲突, 先请 @copilot 帮忙解决, 随后以 Merge branch 'main' into kunshang/remove_pip_in_test 提交收尾——这也意味着 xpu.in 中 cohere/cohere_melody、opentelemetry 等部分条目可能混入了 main 的合并内容, diff 语义不够纯粹。
- CI 触发依赖未覆盖新增的 requirements/test/xpu.in(.txt) (testing): 未修复。PR 只有 remove pip install 与 merge main 两个 commit, 评论无回复, 评审人 bigPYJ1151 直接 approve 后合入, 缺口被带到 main。
 - fork 来源 PR 的自动评审被禁用 (other): 维持禁用状态, 没有触发额外一次性评审; 评审覆盖度依赖人工 reviewer。
 - 合并冲突与 CI 触发流程 (other): 通过合并 main 解决冲突, CI 以最终 commit 4bbf81b 跑通并合入; 但合并可能把 main 的依赖条目带入 xpu.in, 使 diff 归因不够纯粹。

风险与影响

- 风险:

1. CI 触发覆盖缺口（已确认，未修复）：`.buildkite/intel_jobs/*.yaml` 中 `source_file_dependencies` 只列 `vllm/`、`tests/...` 与自身 `yaml`，不含 `requirements/test/xpu.in(.txt)`；今后任何只改 XPU 依赖文件的 PR 都不会触发对应 XPU 测试，依赖回归只能在合入 `main` 之后才暴露。CodeRabbit 还点名了 `.buildkite/ci_config_intel.yaml` 需要同步。
2. 依赖安装从“步骤级联网安装”变成“镜像级预装”，失败面转移：删除 `pip install tensorizer==2.10.1`、`runai-model-streamer[...]`、`av`、`matplotlib`、`tblib` 等命令后，这些包必须由 XPU 镜像（`image_build_xpu.sh`）在构建阶段提供；一旦镜像构建缓存未刷新或构建失败，`model_executor / multi-modal / example` 等整组步骤会因缺包直接报错，而不是像以前那样在步骤内临时装齐。
3. 锁文件级别的版本变化可能带来隐性回归：`gpt-oss 0.0.9` → `0.0.5`（连带移除 `jupyter-client/jupyter-core`），`numpy 2.3.5` → `2.2.6`，`open-clip-torch` 从 `--no-deps` 安装变为参与完整依赖解析（新增 `ftfy` 等传递依赖）。这些变化没有在 PR 描述中给出任何验证结论，PR body 的 Test Plan / Test Result 章节是空的。
4. 可复现性的双刃剑：`av` 等包从“安装当日最新版”变为锁定的 `av==18.1.0`，有利于复现但意味着行为变化点从 CI 运行时悄悄前移到了镜像构建时；镜像 tag 与 `requirements` 的对应关系没有在本 PR 中显式说明。

- 影响:

1. 对用户：不改变任何推理路径与 API 行为，最终用户无直接感知；间接影响是 XPU 平台若出现依赖类回归，可能更早（或更晚，视触发覆盖而定）被 CI 捕获。
2. 对系统 / CI：XPU 测试步骤的命令串显著缩短（去掉联网 `pip install`），理论上每个 step 启动更快、失败定位更清晰（依赖缺失会集中在镜像阶段）；但 CI 的失败模式从“步骤内安装失败”集中为“镜像构建失败”，且镜像构建是 XPU 全组的 `depends_on`，失败时会阻塞整个 Intel 流水线。
3. 对团队：明确了一条新的约束——XPU 测试新增依赖应写进 `requirements/test/xpu.in` 而不是 `yaml` 内的 `pip install`；Intel CI 维护者需要额外留意 `requirements` 变更与镜像重建的联动，同时需要补上触发依赖（跟进 CodeRabbit 的建议）。
4. 对仓库依赖管理：`xpu.in/xpu.txt` 与 `cuda.txt`、`tpu.txt` 等并列，成为 XPU 平台单一依赖真源，后续依赖升级类 PR（如 `tpu-inference` 升版、`CUTLASS DSL` 升版）的同类模式可能继续扩展到 XPU。- 风险标记：CI 触发覆盖缺口，依赖安装改为镜像预装，锁文件版本回退（`gpt-oss 0.0.9`→`0.0.5`、`numpy 2.3.5`→`2.2.6`），`open-clip-torch` 由 `--no-deps` 改为完整解析，PR 描述缺少测试结果

关联脉络

- PR #56463 [XPU][CI] skip DeepSeek-V4.1-Flash in test_tensor_schema.py: 同样修改 `.buildkite/intel_jobs/models_multimodal_intel.yaml` 的 Multi-Modal Processor 步骤；本 PR head 中该步骤仍保留 `--deselect=DeepSeek-V4.1-Flash` 这一行，只是删掉了前面的 `pip install av matplotlib ftfy` 与 `open-clip-torch --no-deps`，属于同一 job 定义的连续演

进。

- PR #56329 [CI] Give every Buildkite step an explicit key, and a hook to enforce it: 同一批 .buildkite/intel_jobs/*.yaml (含 test-intel.yaml、misc_intel.yaml、models_multimodal_intel.yaml) 刚在 #56329 中被要求补齐显式 step key 并由 pre-commit 钩子强制；本 PR 继续在这些 yaml 上做命令裁剪，必须保持 step key 不变，属于同一条 CI 配置维护线。