

PR #45654 完整报告

vllm-project/vllm

[CI/Build] Avoid duplicate ViT CG test introduced by accident

合并时间: 2026-06-17 20:23

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/45654>

执行摘要

- 一句话: 移除重复的 ViT CUDA Graph 测试用例
- 推荐动作: 可直接合并。该 PR 是纯粹的 CI 优化, 逻辑清晰, 适合快速批准。

功能与动机

PR body 指出: "Currently, ViT CG tests run on both **Multi-Modal Models (Standard) 2: qwen3 + gemma** and **Multi-Modal Models (Standard) 4: other + whisper**, which waste CI resource." 因此移除 qwen+gemma 套件中的重复用例。

实现拆解

1. 删除重复测试: 在 Multi-Modal Models (Standard) 2: qwen3 + gemma 的 commands 中移除 `pytest -v -s models/multimodal/generation/test_vit_cudagraph.py -m core_model` 这一行。
2. 调整测试归属: 在 Multi-Modal Models (Standard) 4: other + whisper 的 commands 中, 将原本通过 `--ignore` 排除的 `test_vit_cudagraph.py` 从忽略列表中移除, 并显式添加 `pytest -v -s models/multimodal/generation/test_vit_cudagraph.py -m core_model` 命令。
3. AMD 镜像同步: 为 Standard 4 套件添加 AMD mirror 配置, 设置 `soft_fail: true`、`mi325_1` 设备和 `image-build-amd` 依赖。此调整基于 AndreasKaratzas 的评论, 他表示目前 AMD 上有 `test_phi3v.py` 失败, 先标记为 `soft_fail` 以保持 CI 绿色。
4. 仅涉及 CI 配置: 变更仅限 `.buildkite/test_areas/models_multimodal.yaml` 文件, 无代码逻辑改动。

关键文件:

- `.buildkite/test_areas/models_multimodal.yaml` (模块 CI 配置; 类别 config; 类型 configuration): 核心变更文件, 通过删除和添加测试命令消除重复 ViT CUDA Graph 测试, 并增加 AMD mirror 配置。

关键符号: 未识别

评论区精华

讨论集中在 AMD CI 的适配: AndreasKaratzas 指出 `test_phi3v.py` 在 AMD 上失败, 建议添加 `soft_fail: true`, 并承诺后续修复。PR 作者采纳了建议。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。仅 CI 配置变更，测试逻辑无修改。唯一风险是 AMD mirror 新增 soft_fail 后可能掩盖真正的失败，但作者已计划后续修复。
- 影响：对系统性能无影响。CI 资源消耗降低：每次运行减少一次 test_vit_cudagraph.py 的执行。团队协作上，AMD CI 配置的更新为后续维护提供了明确入口。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR