

PR #54957 完整报告

vllm-project/vllm

[Bugfix][CI] Set cudagraph_mode=FULL for the Ernie4.5-VL ViT cudagraph test

合并时间: 2026-09-03 02:18

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/54957>

执行摘要

- 一句话: 修复 Ernie4.5-VL ViT cudagraph 测试因模型未编译导致的 CI 失败。
- 推荐动作:

功能与动机

根据 PR body, 测试 `test_vit_cudagraph_image[ernie45_vl]` 在 #54782 后失败, 错误信息为 'piecewise CUDA graphs (cudagraph_mode=FULL_AND_PIECEWISE) unavailable, model is not torch-compiled'. 根本原因是模型 `Ernie4_5_VLMoeModel` 的 `@support_torch_compile` 装饰器被注释掉, 导致模型未编译, piecewise cudagraphs 无法分区, 测试请求默认模式触发了新 guard。

实现拆解

1. 问题识别: 在 CI 中, Ernie4.5-VL ViT cudagraph 测试在 MI300 和 MI355 上失败, 错误指向 `cudagraph_utils.py` 中的 guard。这影响了多模态模型测试的可靠性。
2. 原因分析: 检查模型定义 `vllm/models/deepseek_v4/...` 或类似文件, 发现 `Ernie4_5_VLMoeModel` 的 `@support_torch_compile` 装饰器被注释, 因为其文本和视觉专家分离破坏了编译。因此, 模型未 torch-compiled, piecewise cudagraphs 无法工作。
3. 修复方案: 修改测试文件 `tests/models/multimodal/generation/test_vit_cudagraph.py`, 在 `ernie45_vl` 配置中添加 ``compilation_config_overrides={`

关键文件:

- `tests/models/multimodal/generation/test_vit_cudagraph.py` (模块 `vit/cudagraph`; 类别 `test`; 类型 `test-coverage`): 测试配套; 包含 测试覆盖调整、配置键调整; +7/-0

关键符号: 未识别

评论区精华

- 暂无高价值评论线程

风险与影响

- 风险:
- 影响:

- 风险标记：暂无

关联脉络

- 暂无明显关联 PR