

PR #55266 完整报告

vllm-project/vllm

[XPU][UT] skip GLM-5.3-Flash test on XPU

合并时间: 2026-09-04 14:45

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/55266>

执行摘要

- 一句话: 在多模态模型测试中为 GLM-5.3-Flash 添加 XPU 平台跳过逻辑, 确保 CI 通过。
- 推荐动作: 该 PR 是一个简单、低风险的测试维护操作, 旨在恢复 CI 的绿色状态。虽然改动微小, 但其背后的跨平台支持决策和审核过程值得关注。对于普通工程师, 无需精读; 对于负责 Intel XPU 平台或 CI 管道的工程师, 可快速确认跳过逻辑是否符合预期。

功能与动机

根据 PR body 中引用的 CI 链接, GLM-5.3-Flash 模型在 XPU 上运行时会报出 “Unsupported model on XPU” 错误, 导致集成测试失败。为了保持 CI 绿色状态并明确该模型在当前硬件上的支持状态, 需要在测试中将其跳过。

实现拆解

变更仅涉及一个测试文件, 实现步骤非常直接:

1. 变更入口与目标: 修改 `tests/models/multimodal/processing/test_tensor_schema.py` 中的 `test_model_tensor_schema` 测试函数。
2. 核心逻辑改造: 在现有针对 DeepSeek-V4 和 Kimi-K2.5 模型的平台条件跳过逻辑之后, 新增一个针对 `zai-org/GLM-5.3-Flash` 模型在 `current_platform.is_xpu()` 条件下的 `pytest.skip` 调用。这确保了当测试参数为该模型且运行在 XPU 平台时, 测试会被标记为跳过而非失败。
3. 测试配套: 该变更本身就是对现有测试覆盖的调整, 无需额外的测试配套改动。其影响是确保了 Intel XPU CI 管道在该模型上的稳定性。

关键文件:

- `tests/models/multimodal/processing/test_tensor_schema.py` (模块 测试用例; 类别 `test`; 类型 `test-coverage`): 本次变更的唯一文件, 包含了跳过 GLM-5.3-Flash 在 XPU 上测试的核心逻辑。

关键符号: `test_model_tensor_schema`

关键源码片段

[tests/models/multimodal/processing/test_tensor_schema.py](#)

本次变更的唯一文件, 包含了跳过 GLM-5.3-Flash 在 XPU 上测试的核心逻辑。

```
# 以下片段展示了 test_model_tensor_schema 函数中新增的跳过逻辑。
# 它插入在已有的针对 Kimi-K2.5 和 DeepSeek-V4 的平台检查之后，
# 确保测试在正确的平台条件下执行。
def test_model_tensor_schema(model_id: str):
    if model_id == "moonshotai/Kimi-K2.5":
        pytest.skip(
            "Kimi-K2.5's offline inference has issues about vision chunks. Fix later."
        )

    if model_id == "deepseek-ai/DeepSeek-V4-Flash-Vision-Exp" and not (
        current_platform.is_cuda()
    ):
        pytest.skip("Deepseek V4 is only supported on CUDA")

    # 新增：在 XPU 平台上跳过不支持的 GLM-5.3-Flash 模型
    if model_id == "zai-org/GLM-5.3-Flash" and (current_platform.is_xpu()):
        pytest.skip("GLM-5.3-Flash is not supported on XPU")

    model_info = HF_EXAMPLE_MODELS.find_hf_info(model_id)
    # ... 后续测试逻辑不变 ...
```

评论区精华

核心讨论围绕一个潜在风险展开：

- 争议点：审核者 jikunshang 最初提出了一个担忧：“oh sorry, this may break AMD?”，即担心在条件判断中使用 `current_platform.is_xpu()` 可能会意外影响 AMD 平台上的测试执行。
- 决策结论：作者 mayuyuace 确认已经检查并处理了这个问题（可能通过调整条件逻辑），随后 jikunshang 审批通过。这表明团队在跨平台兼容性上保持了警惕。
- 未解决疑虑：无。讨论在作者澄清后迅速结束。
 - 跨平台测试跳过逻辑的兼容性担忧 (correctness): 作者 mayuyuace 确认已检查并确保逻辑正确，不会影响 AMD。审核者随后批准。

风险与影响

- 风险：
 1. 回归风险：极低。改动仅限于在特定条件下跳过测试，不影响任何产品代码逻辑、数据流或现有功能。最坏的情况是意外跳过了其他平台上的必要测试，但经过审核确认，此风险已排除。
 2. 兼容性风险：本次改动本身就是为解决 XPU 兼容性问题。需关注未来是否会有其他测试用例因类似原因失败。
- 影响：
 1. 对用户的影响：无直接影响。用户不会感知到此变更。
 2. 对系统的影响：无。不改变系统行为、性能或 API。

3. 对团队的影响：为 Intel XPU CI 团队消除了一处持续性的测试失败，提升了 CI 信号的清晰度和稳定性。同时，明确了 GLM-5.3-Flash 模型在 XPU 上当前不被支持的状态。
- 风险标记：测试覆盖调整

关联脉络

- PR #53037 [XPU] Fix device assignment for DP external LB: 同属 Intel XPU 平台适配与修复类别，致力于确保 vLLM 在 XPU 上的正确运行。
- PR #55245 [SpecDecode]Fix spec decode warmup device selection: 同样是修复与设备 / 平台（CUDA/XPU）硬编码或选择逻辑相关的错误，确保功能在不同硬件上可用。