PR #42923 完整报告

vllm-project/vllm

Revert checkpoint specific workaround in Transformers modelling backend

合并时间：2026-05-18 16:31

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42923

执行摘要

- 一句话：回退针对 Gemma3 特殊权重的 hack
- 推荐动作：该 PR 是小型清理，不值得精读。但可作为建模后端维护的参考：避免在通用路径中放置特定模型的临时 hack。

功能与动机

PR body 明确说明：'This kind of fix does not belong in in the Transformers modelling backend.' 同时，review 中（#3256477280）hmellor 指出该 hack 并非 CI 失败的真实原因，且 Gemma3 测试在 AMD nightly 中通过。

实现拆解

1. 定位待删除代码：在文件 vllm/model_executor/models/transformers/base.py 的 _create_hf_to_vllm_mapper 方法中，找到无条件为 Gemma3 添加的 regex 映射（约第 356-362 行）。
2. 删除特设映射：移除 vision_tower 属性检查和对应的正则替换逻辑，该逻辑用于处理旧版 Transformers 保存的权重中多余的 vision_model 层级。
3. 保留通用逻辑：其他通用权重重命名、前缀标准化、键忽略等逻辑保持不变。
4. 无测试变化：本次仅删除 8 行源码，无新增测试。

关键文件：

- vllm/model_executor/models/transformers/base.py（模块 建模后端；类别 source；类型 data-contract；符号 _create_hf_to_vllm_mapper）：删除 Gemma3 特定的 hack，恢复通用权重重映射逻辑。

关键符号：_create_hf_to_vllm_mapper

关键源码片段

vllm/model_executor/models/transformers/base.py

删除 Gemma3 特定的 hack，恢复通用权重重映射逻辑。

```
# 变更集中在 _create_hf_to_vllm_mapper 方法中
# 删除以下代码块（base 版本有，head 版本已移除）：
# Gemma3 checkpoints saved with older Transformers versions include an
```

```
# extra `vision_model` level that the current AutoModel no longer has.
vision_tower = getattr(self.model, "vision_tower", None)
if vision_tower is not None and not hasattr(vision_tower, "vision_model"):
    orig_to_new_regex[
        re.compile(r"^(?:model\.)?vision_tower\.vision_model\.(.+)")
    ] = r"model.vision_tower.\1"
```

评论区精华

AndreasKaratzas 提议用更通用的方式处理（基于 transformers 版本检查），但 hmellor 线下讨论后确认：

- 该 hack 并非 AMD CI 失败的原因（Gemma3 测试在 AMD nightly CI 中通过）。
- AMD CI 使用的 checkpoint 并非旧版。
- 故直接 revert，并计划通过 CI 验证。最终 AMD CI 通过后合并。
- 替代方案讨论 (design): hmellor 线下确认该 hack 并非 CI 失败原因，决定直接删除，并运行 CI 验证

风险与影响

- 风险：低风险。删除的 hack 仅在特定旧版 Gemma3 checkpoints 下生效，若用户仍使用此类 checkpoint，可能因缺少该映射导致权重加载失败。但由于 Transformers 已更新至新版（≥5.0），且 PR#42909 已被回退，实际影响极小。
- 影响：影响范围：仅影响使用旧版 Transformers 保存的 Gemma3 checkpoint 加载场景。对于大多数用户无影响。CI 已通过验证。
- 风险标记：低风险清理

关联脉络

- PR #42909 [ROCm][CI] Stabilize ROCm pooling and multimodal CI: 本 PR 回退了该 PR 中引入的 Gemma3 hack