PR #42673 完整报告

vllm-project/vllm

[Model Runner v2] Support reload weights (sleep mode)

合并时间：2026-05-16 00:41

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42673

执行摘要

- 一句话：MR v2 支持权重重载（sleep mode）
- 推荐动作：建议在完成 v2 完全迁移后，移除此委托方法并直接内联实现。同时应补充单元测试覆盖 reload_weights 调用后的缓存重置行为。

功能与动机

作为从 Model Runner v1 迁移到 v2（issue #41286）的一部分，原先使用 VLLM_USE_V2_MODEL_RUNNER=1 运行 test_deep_sleep 时会因 GPUModelRunner 缺少 reload_weights 属性而抛出 AttributeError。本 PR 新增该方法以消除该错误。

实现拆解

1. 在 vllm/v1/worker/gpu/model_runner.py 的 GPUModelRunner 类中添加 reload_weights 方法，位于 get_model 之后。
2. 方法内部通过延迟导入 GPUModelRunnerV1 并调用其 reload_weights，实现权重重载逻辑的复用。这符合渐进式迁移策略：在 v2 完全就绪前，先委托给成熟的 v1 实现。
3. 在权重重载后，显式调用 self.reset_encoder_cache() 和 self.reset_mm_cache()，确保编码器和多模态缓存失效，避免模型使用旧权重计算出的嵌入。
4. 没有新增测试文件，但 PR body 中报告了 test_deep_sleep 测试通过。

关键文件：

- vllm/v1/worker/gpu/model_runner.py（模块 运行器；类别 source；类型 data-contract；符号 reload_weights）：核心变更文件，新增 reload_weights 方法，修复 v2 model runner 缺少该接口的问题。

关键符号：GPUModelRunner.reload_weights

关键源码片段

vllm/v1/worker/gpu/model_runner.py

核心变更文件，新增 reload_weights 方法，修复 v2 model runner 缺少该接口的问题。

```
def reload_weights(self, *args, **kwargs) -> None:
    # TODO(Wentao): Use full version instead of import when fully migrated to v2
    from vllm.v1.worker.gpu_model_runner import GPUModelRunner as GPUModelRunnerV1
```

```
GPUModelRunnerV1.reload_weights(self, *args, **kwargs) # type: ignore[arg-type]
self.reset_encoder_cache()
self.reset_mm_cache()
```

评论区精华

gemini-code-assist[bot] 在 review 中指出：权重重载后，之前缓存的编码器 / 多模态嵌入是基于旧权重计算的，会变得陈旧。建议在 reload_weights 末尾添加 reset_encoder_cache() 和 reset_mm_cache() 调用。该建议被采纳并体现在最终提交中。

- 权重重载后需要重置编码器和多模态缓存 (correctness): 建议被采纳，最终提交加入了这两个重置调用。

风险与影响

- 风险：低风险。新增方法仅 8 行，且完全委托给 v1 实现，逻辑上不会引入新错误。但依赖方（如 sleep mode 测试）需要确保在调用权重重载后正确处理缓存重置，本 PR 已自动处理。
- 影响：影响范围：仅影响使用 VLLM_USE_V2_MODEL_RUNNER=1 且需要权重重载的场景（如 sleep mode）。影响程度：修复了阻塞测试的 bug，是 v2 迁移的必要步骤。
- 风险标记：缺少测试覆盖

关联脉络

- PR #41286 [Feature]: Migration from Model Runner v1 to Model Runner v2: 本 PR 是 issue #41286 中列出的子任务之一，推进 Model Runner v1 到 v2 的迁移。