PR #6932 完整报告

verl-project/verl

[rollout] fix: disable vLLM multimodal processor cache for vLLM < v0.22.0

合并时间：2026-07-13 14:54

原文链接：http://prhub.com.cn/verl-project/verl/pull/6932

执行摘要

- 一句话：禁用 vLLM 多模态处理器缓存修复
- 推荐动作：值得精读。该 PR 展示了一个典型的上游兼容性修复模式：在引擎初始化前通过版本检测注入 workaround，同时保留用户覆写路径。适合作为多模态训练中缓存问题处理的参考。

功能与动机

vLLM 在低于 0.22.0 版本时，多模态处理器缓存会在 pause/resume 过程中反序列化，导致 AssertionError: Expected a cached item for mm_hash='...' 错误。该问题在 verl 调用 update_weights() 后触发 pause_generation(clear_cache=True) 时出现。参见 vLLM issue #42995 及 PR #43001。

实现拆解

1. 修改 _preprocess_engine_kwargs 方法：在 verl/workers/rollout/vllm_rollout/vllm_async_server.py 中，将原本的 no-op 实现改为版本检测逻辑。
2. 版本条件判断：通过比较 _VLLM_VERSION 与 0.22.0，若低于该版本则执行缓存禁用。
3. 设置引擎参数：使用 engine_kwargs.setdefault("mm_processor_cache_gb", 0)，确保仅当用户未显式指定时才设为 0，保留用户覆盖能力。
4. 注释更新：更新方法 docstring 并添加引用链接指向上游 PR。

关键文件：

- verl/workers/rollout/vllm_rollout/vllm_async_server.py（模块 rollout；类别 source；类型 core-logic；符号 _preprocess_engine_kwargs）：核心修改文件：在 _preprocess_engine_kwargs 方法中增加版本判断，当 vLLM < 0.22.0 时自动禁用多模态处理器缓存。

关键符号：_preprocess_engine_kwargs

关键源码片段

verl/workers/rollout/vllm_rollout/vllm_async_server.py

核心修改文件：在 _preprocess_engine_kwargs 方法中增加版本判断，当 vLLM < 0.22.0 时自动禁用多模态处理器缓存。

```
# verl/workers/rollout/vllm_rollout/vllm_async_server.py

# 在文件顶部有版本导入
from packaging import version
from vllm import VllmVersion as _VLLM_VERSION

def _preprocess_engine_kwargs(self, engine_kwargs: dict) -> None:
    """Mutate engine_kwargs in-place before the CLI args dict is built."""
    if _VLLM_VERSION < version.parse("0.22.0"):
        # Work around multimodal processor cache desync across pause/resume.
        # 当 vLLM 版本低于 0.22.0 时，多模态处理器缓存会在 pause/resume
        # 过程中反序列化，导致 AssertionError。详见上游 PR #43001。
        # setdefault 确保用户显式设置不被覆盖。
        engine_kwargs.setdefault("mm_processor_cache_gb", 0)
```

评论区精华

Review 中主要讨论了三点：

1. 实现位置：@wuxibin89 建议将逻辑从 rollout.py 配置层移到 vllm_async_server.py，作者随后采纳。
2. Shell 脚本问题：@gemini-code-assist[bot] 指出示例 shell 脚本中有 bash 语法错误（反斜杠引号问题），但该脚本最终被排除在 PR 外。
3. 影响范围：@tardis-key 询问是否仅为 NPU 问题，作者确认上游 vLLM issue 同时影响 GPU 和 NPU。

- 实现位置选择：配置层 vs 引擎层 (design): 逻辑移至 vllm_async_server.py 的 _preprocess_engine_kwargs 方法中。
- 是否仅 NPU 受影响 (question): 该问题是 vLLM 通用 bug，跨硬件平台。
- Shell 脚本语法问题 (style): 相关脚本最终被排除在 PR 外，仅保留核心源码修改。

风险与影响

- 风险：风险极低：
  - 仅修改一个方法，逻辑简单，且使用 setdefault 确保用户显式设置不被覆盖。
  - 不影响 vLLM >= 0.22.0 的用户。
  - 该缓存禁用可能会轻微增加多模态处理开销（每次重新构建处理器），但避免崩溃更重要。
- 影响：影响范围有限：
  - 仅影响使用 vLLM < 0.22.0 且运行多模态模型（如 Qwen3-VL）的用户。
  - 对训练正确性无负面影响，仅避免断言失败。
  - 对性能有轻微潜在影响（禁用缓存后每次 pause/resume 重新构建处理器），但通常可忽略。
  - 风险标记：上游依赖

关联脉络

- PR #6937 fix CI issue: 作者提到失败 CI 由 PR#6937 修复，该 PR 是 PR#6932 的依赖。
- PR #6982 [trainer,doc] fix: workaround vLLM multimodal cache in verl 0.8.0, add Qwen3.5 397B: 同一功能线：PR#6982 也处理了 vLLM 多模态缓存问题，不过是在更高版本中通过不同方式规避。