PR #5961 完整报告

verl-project/verl

[rollout, vllm] fix: auto-convert disable_mm_preprocessor_cache to mm_processor_cache_gb for vllm >= 0.13.0

合并时间：2026-04-14 14:26

原文链接：http://prhub.com.cn/verl-project/verl/pull/5961

执行摘要

本 PR 自动处理 vLLM 0.13.0 中移除的多模态预处理器缓存参数，通过版本检查逻辑将 disable_mm_preprocessor_cache 转换为 mm_processor_cache_gb，确保多模态训练脚本在升级后无需修改即可正常运行，修复了因参数废弃导致的训练崩溃问题。

功能与动机

根据 Issue #5959 报告，vLLM 0.13.0 移除了 disable_mm_preprocessor_cache 参数，导致使用该参数的多模态训练脚本（如 Qwen3-VL-8B-Instruct）失败并抛出 ActorDiedError。PR 旨在自动转换此参数，维持向后兼容性，使用户在升级 vLLM 版本时无需手动调整配置。

实现拆解

核心改动在 verl/workers/rollout/vllm_rollout/vllm_async_server.py 的 launch_server() 函数中：

```
if "disable_mm_preprocessor_cache" in engine_kwargs:
    if _VLLM_VERSION >= version.parse("0.13.0"):
        disable_cache = engine_kwargs.pop("disable_mm_preprocessor_cache")
        if disable_cache:
            if "mm_processor_cache_gb" not in engine_kwargs:
                engine_kwargs["mm_processor_cache_gb"] = 0
```

此外，更新了 16 个示例脚本（如 examples/grpo_trainer/run_qwen3_vl-8b_npu.sh），将 disable_mm_preprocessor_cache=True 直接替换为 mm_processor_cache_gb=0，确保脚本一致性。

评论区精华

gemini-code-assist[bot] 指出初始实现存在逻辑缺陷：

> " 当 disable_mm_preprocessor_cache 为 True 时，代码记录为'自动转换为 mm_processor_cache_gb=0'，但如果用户已显式设置 mm_processor_cache_gb 为非零值，setdefault 不会覆盖它，导致误导性日志和潜在错误配置。" 讨论后，代码被调整以更稳健地处理参数冲突，并移除了警告日志。

风险与影响

风险：版本检查准确性依赖 _VLLM_VERSION 变量；参数优先级处理在极端场景下可能未覆盖；缺少 CI 测试覆盖，因需多模态环境。 影响：用户无需修改脚本即可兼容新版本 vLLM，系统修复了训练崩溃，团队降低了维护负担。

关联脉络

此 PR 反映了处理外部依赖 breaking change 的常见模式。近期历史 PR 中，如 #5934 涉及 vLLM 性能优化，但未直接处理参数废弃；整体上，verl 项目持续适配 vLLM 版本升级，确保多模态训练稳定性。