

PR #48014 完整报告

vllm-project/vllm

[Core] Move MRV1 `late\_interaction\_runner.py` out of MRV2 subtree

合并时间: 2026-07-09 02:30

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/48014>

执行摘要

将 `late_interaction_runner.py` 从 MRV2 子树 `vllm/v1/worker/gpu/pool/` 移动到 MRV1 核心模块 `vllm/v1/pool/`，并更新所有导入引用。纯代码结构清理，无功能变更。

功能与动机

该文件在 PR #36159 中被引入，但错误地放置在 `vllm/v1/worker/gpu/pool/`（MRV2 的源码子树）下。实际上它与 `pool` 模块中的其他文件（如 `metadata.py`）同属 MRV1 核心，应放在 `vllm/v1/pool/` 下。本次移动修复了文件位置错误，使代码组织与模块职责一致。

实现拆解

1. 重命名文件：将 `vllm/v1/worker/gpu/pool/late_interaction_runner.py` 移动到 `vllm/v1/pool/late_interaction_runner.py`，内容完全不变。
2. 更新 `gpu_model_runner.py` 中的导入：将 `from vllm.v1.worker.gpu.pool.late_interaction_runner import LateInteractionRunner` 替换为 `from vllm.v1.pool.late_interaction_runner import LateInteractionRunner`。
3. 更新测试文件导入：同步更新 `tests/v1/worker/test_late_interaction_runner.py` 中的导入路径。

无代码逻辑变更，仅导入路径调整。

评论区精华

无实质性讨论，WoosukKwon 直接批准。

风险与影响

风险极低，仅涉及文件路径和导入语句调整，无任何运行时行为改变。影响范围仅限于直接导入旧路径的两个文件。

关联脉络

本 PR 是对 PR #36159（引入 late interaction 功能）的修正，修复了文件放置位置错误。属于代码结构清理的一部分。