

# PR #40759 完整报告

vllm-project/vllm

[Examples] Resettle Disaggregated examples.

合并时间: 2026-05-06 16:20

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/40759>

## 执行摘要

本 PR 按 RFC #29362 的建议, 将分布在各处的 disaggregated 示例统一集中到 [examples/disaggregated/](#) 目录下, 同时更新了文档和脚本中的引用路径, 使用户能更直观地按使用场景查找示例。变更仅涉及文件搬迁和路径修正, 无逻辑改动。

## 功能与动机

Issue #29362 指出当前 examples 第一级按 offline\_inference / online\_serving 组织, 用户需在多个文件夹中查找特定用途示例 (例如disaggregatedprefill的示例散落在多个子目录中)。提议按场景作为第一级目录。本 PR 率先迁移 disaggregated 相关示例, 验证目录结构可行性并方便后续其他场景 (如 pooling、structured\_outputs) 的迁移。

## 实现拆解

1. 在 examples/ 下新建 disaggregated/ 目录, 按二级功能划分子目录: disaggregated\_serving、disaggregated\_encoder、flexkv\_connector、lmcache、example\_connector、kv\_load\_failure\_recovery\_offline、mooncake\_connector 等。
2. 使用 git mv 将 20+ 个文件从旧位置移动到新位置, 保持文件内容不变。
3. 更新 docs/features/disagg\_prefill.md 中的链接标签和描述文字, 修正 review 中指出的两处问题。
4. 对移动后的脚本内部 (如 argparse help 中的示例命令) 同步更新路径。
5. 无测试或配置变更, 属于纯组织工作。

(本 PR 不涉及逻辑变更, 无必要展示源码片段。)

## 评论区精华

- gemini-code-assist发现 docs/features/disagg\_prefill.md 中链接标签缺少 examples/ 前缀 (high priority), 且“MooncakeConnector”段落错误地沿用了“ExampleConnector”的描述 (copy-paste 错误)。作者通过两个独立 commit 修复了这两处问题, 最终获得 maintainer DarkLight1337 的 approval。

## 风险与影响

- 风险: 极低, 仅可能存在未及时更新的路径引用导致用户无法找到示例的风险。本次 PR 已全面清理, 并由 reviewer 确认。

- 影响：用户需更新运行示例的命令路径；开发者维护时能更快速定位 disaggregated 相关示例。

## 关联脉络

- 关联 Issue #29362 (RFC: Resettle examples) ，本 PR 是首个针对 disaggregated 场景的落实。未来可能陆续迁移 pooling、structured\_outputs 等其他场景的示例。
- PR 内注释提及 PDController (#15343) 未来可能取代部分 demo 功能。