PR #38628 完整报告

vllm-project/vllm

[Docs] PD with Nixl compat matrix

合并时间：2026-03-31 23:01

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38628

执行摘要

新增 NixlConnector 兼容性矩阵文档，澄清分散预填充功能支持，提升文档完整性以指导用户配置，属于常规文档维护。

功能与动机

本 PR 旨在澄清在 NixlConnector 下分散设置中支持的功能列表。根据 PR body，这是首次尝试，重点是内容而非 UX，以帮助用户理解兼容性，避免配置错误。PR body 中表述：'First tentative in clarifying the list of supported features in disaggregated setup with NixlConnector. Focusing on content here rather than UX'。

实现拆解

主要变更包括三个文档文件：

- 新增 docs/features/nixl_connector_compatibility.md：包含详细的兼容性矩阵，使用表格列出不同模型类型（如 Dense Transformers、MLA、MoE）对各项功能（如基本 PD、Spec Decode、Hetero TP）的支持状态（✅/🟠/❌等），并附加脚注说明限制条件。
- 修改 docs/features/disagg_prefill.md：在 NixlConnector 部分添加链接：'For feature compatibility details, see NixlConnector Compatibility Matrix.'。
- 修改 docs/features/nixl_connector_usage.md：在开头添加引用：'For feature compatibility details (supported model architectures, TP configurations, and feature interactions), see the NixlConnector Compatibility Matrix.'。

无代码逻辑变更，纯文档更新。

评论区精华

Review 讨论较少：

- gemini-code-assist[bot] 评论：'I have no feedback to provide as there were no review comments to assess.'
- robertgshaw2-redhat 直接批准。 无争议点或深度技术讨论，PR 顺利合并。

风险与影响

风险：主要涉及文档准确性，兼容性矩阵需随代码变更及时更新，否则可能导致用户误解支持状态；无代码回归、性能或安全风险。 影响：限于文档用户，提供清晰的兼容性指南，有助于

减少配置错误和提升用户体验；无系统层面影响。

关联脉络

与历史 PR 相关：

- PR 36742：更新 EPD 脚本参数文档，同样涉及 kv-connector 和文档改进，标签包括 'documentation' 和 'kv-connector'，显示文档工作的延续性。
- PR 38554：修复 kv_offload 相关 bug，标签有 'kv-connector'，反映 kv-connector 模块的活跃开发，本 PR 的文档更新与之协同，确保用户文档与实际功能同步。 整体趋势：vLLM 项目在 kv-connector 领域持续优化，文档更新跟进功能演进。