PR #42342 完整报告

vllm-project/vllm

[Bug] Fix DeepSeek V4 `AttributeError: module 'cutlass.cute.nvgpu' has no attribute 'LoadCacheMode'`

合并时间：2026-05-14 17:00

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42342

执行摘要

- 一句话：修复 DeepSeek V4 因 cutlass 版本 API 变动导致的崩溃
- 推荐动作：可快速合并，无必要精读。注意跟进 @ZJY0516 在 #42438 中的改动，确保版本管理一致性。

功能与动机

PR body 中给出的运行错误显示 AttributeError: module 'cutlass.cute.nvgpu' has no attribute 'LoadCacheMode'，该错误发生在 dequant_gather_k_cutedsl.py 第 106 行调用 cpasync.CopyG2SOp(cute.nvgpu.LoadCacheMode.GLOBAL) 时，原因是新版本 cutlass DSL 已弃用 nvgpu.LoadCacheMode。

实现拆解

1. 定位根因：CuteDSL 内核 dequant_gather_k_cutedsl.py 使用了已在 nvidia-cutlass-dsl>=4.5.0 中移除的 cutlass.cute.nvgpu.LoadCacheMode。
2. 修改依赖：在 requirements/cuda.txt 中将版本约束从 >=4.4.2 收紧为 ==4.5.0，确保始终使用兼容的 API。
3. 仅改一行：变更仅涉及 requirements/cuda.txt 中的版本号。

关键文件：

- requirements/cuda.txt（模块 依赖配置；类别 infra；类型 configuration）：唯一修改的文件，将 nvidia-cutlass-dsl 版本从 >=4.4.2 固定为 ==4.5.0，直接修复错误。

关键符号：未识别

评论区精华

原作者 @gau-nernst 指出他开发时使用 nvidia-cutlass-dsl==4.5.0，当时 cpasync.LoadCacheMode 已弃用，建议固定版本。@mgoin 担心未来 breaking change，提议设置上限或精确固定。最终采用精确固定到 4.5.0。@ZJY0516 告知已在另一个 PR #42438 中将依赖改为 nvidia-cutlass-dsl[cu13]。

- 依赖版本是否应固定或设上限 (design): 固定到 4.5.0 版本，确保与当前 CuteDSL 内核兼容。

风险与影响

- 风险：风险低。仅修改依赖版本，不涉及代码逻辑。潜在风险是固定到 4.5.0 可能导致未来因依赖过旧而错过安全或性能更新，但可通过后续版本升级 PR 缓解。
- 影响：影响范围：所有使用 DeepSeek V4 且安装了 nvidia-cutlass-dsl>=4.5.0 的用户。修复后该模型可正常启动。团队需注意后续依赖管理的稳定性。
- 风险标记：依赖版本固定可能阻碍未来更新

关联脉络

- PR #42438 [ci] change dependencies to nvidia-cutlass-dsl[cu13]：同一依赖行被改动，需关注是否冲突或需协调。