PR #42991 完整报告

vllm-project/vllm

[CI/Build] Bump nvidia-cutlass-dsl to 4.5.1

合并时间：2026-05-19 07:58

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42991

执行摘要

- 一句话：bump cutlass-dsl 至 4.5.1 修复 Blackwell GDN ICE
- 推荐动作：建议快速合并。PR 简单、测试充分、风险低，修复了明确的 JIT 编译崩溃问题。

功能与动机

原 pin 的 4.5.0 版本在 Blackwell GPU 上运行 FlashInfer GDN prefill kernel 时触发 JIT compile ICE（内部编译器错误）。升级至 4.5.1 后相关测试全部通过（1563 passed, 0 failed），修复了 GDN 功能的阻碍。

实现拆解

1. 修改依赖版本：在 requirements/cuda.txt 中将 nvidia-cutlass-dsl[cu13]==4.5.0 改为 ==4.5.1。
2. 无其他代码变更：只需调整依赖版本号，无需修改 Python 或 CUDA 源码。

关键文件：

- requirements/cuda.txt（模块 依赖管理；类别 docs；类型 documentation）：唯一变更文件，修改了 nvidia-cutlass-dsl 版本 pin。

关键符号：未识别

评论区精华

无实质性讨论，review 仅为自动机器人和两个 approve。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。仅依赖版本号变化，无 API 或行为变更。如果 4.5.1 引入回归，会导致 GDN 相关测试失败或运行时错误，但 cutlass-dsl 从 4.5.0 到 4.5.1 是小版本修复，兼容性有保障。
- 影响：影响仅限使用 NVIDIA Blackwell GPU（如 GB200）且启用 FlashInfer GDN 预填充内核的用户。修复了之前可能出现的 JIT 编译崩溃，使 GDN 功能正常可用。对非 Blackwell 或其他功能无影响。

- 风险标记：依赖版本号变更，仅影响 Blackwell GDN 场景

关联脉络

- PR #42342 [Bug] Fix DeepSeek V4 AttributeError: module 'cutlass.cute.nvgpu' has no attribute 'LoadCacheMode': 引入 4.5.0 的 pin 的 PR，本 PR 修复了该版本引入的 Blackwell GDN ICE 问题。