PR #25570 完整报告

sgl-project/sglang

Use triton_attn as default vision attention on B300 (SM103)

合并时间：2026-05-19 11:00

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25570

执行摘要

- 一句话：B300 默认视觉注意力回退为 triton_attn
- 推荐动作：此 PR 改动简单直接，适合快速合入以确保 B300 上的多模态功能可用。建议关注后续 FA4 在 B300 上的验证进度，验证通过后恢复 fa4 默认值。

功能与动机

FA4 在 B300 (SM103) 上未经验证，直接使用可能导致错误或性能问题。PR body 明确指出 'FA4 is not yet validated on B300 (SM103)'，因此需要回退到更稳定的 triton_attn 后端。

实现拆解

1. 修改 python/sglang/srt/layers/attention/vision.py 中 _determine_attention_backend 方法的 CUDA Blackwell 分支判断逻辑，将 elif major == 10: 改为 elif major == 10 and minor != 3:，使得 SM103 (B300) 不会进入 fa4 分支，而是走 else 分支使用 triton_attn。
2. 原有逻辑中 major == 10 表示所有 Blackwell 架构 GPU 都使用 fa4，现在排除 SM103，对其余 Blackwell (如 SM100) 仍保留 fa4。
3. 无其他文件修改，无测试配套变更。

关键文件：

- python/sglang/srt/layers/attention/vision.py（模块 视觉注意力；类别 source；类型 core-logic；符号 _determine_attention_backend）：核心改动文件：修改了视觉注意力后端选择逻辑，为 B300 (SM103) 添加了排除条件。

关键符号：_determine_attention_backend

评论区精华

无 review 评论。PR 仅由 mickqian 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：改动仅一行条件判断，且仅影响 B300 硬件的默认后端选择。逻辑清晰，无功能影响。若后续验证通过，需再次更新以恢复 fa4 支持。

- 影响：影响范围很小，仅影响在 B300 (SM103) 上运行多模态模型的用户，其视觉注意力后端将默认使用 triton_attn 而非 fa4，B200/GB200 用户无影响。性能可能略有下降，但保证了正确性。
- 风险标记：无测试覆盖

关联脉络

- 暂无明显关联 PR