

# PR #31062 完整报告

sgl-project/sglang

Revert "[Tiny] Enable Full Cuda Graph with Page size = 1"

合并时间: 2026-07-14 03:23

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/31062>

## 执行摘要

- 一句话: 回退 Full Cuda Graph 与 `page_size=1` 兼容变更
- 推荐动作: 建议快速合并以恢复 CI, 后续若需重新启用该功能, 应充分测试后再提交。

## 功能与动机

PR #30835 的变更破坏了 CI 测试 (<https://github.com/sgl-project/sglang/actions/runs/29276578199/job/86907146007>) , 因此需要回退以恢复 CI 稳定性。

## 实现拆解

1. 回退文件: `python/sglang/srt/arg_groups/overrides.py` 中 `_fa4_page_constraint` 函数被修改。
2. 删除条件: 移除了原先为 full prefill CUDA graph 添加的特殊条件 `and view.cuda_graph_config.prefill.backend != Backend.FULL` 及其注释。
3. 恢复行为: FA4 后端在非 MLA 模型架构上, 如果未使用 MLA 后端且满足其他条件时, 仍会强制将 `page_size` 设为 128。
4. 无其他文件修改或测试配套。

关键文件:

- `python/sglang/srt/arg_groups/overrides.py` (模块 参数覆盖; 类别 source; 类型 core-logic) : 核心变更文件, 回退了 #30835 关于 FA4 与 full prefill CUDA graph 兼容性的逻辑。

关键符号: `_fa4_page_constraint`

## 关键源码片段

`python/sglang/srt/arg_groups/overrides.py`

核心变更文件, 回退了 #30835 关于 FA4 与 full prefill CUDA graph 兼容性的逻辑。

```
# 函数 _fa4_page_constraint 中的关键变更 (删除了 3 行) :
@register_post_process
def _fa4_page_constraint(view: Any) -> dict:
    if (
        (
```

```

        view.attention_backend == "fa4"
        or view.decode_attention_backend == "fa4"
        or view.prefill_attention_backend == "fa4"
    )
    and not view.use_mla_backend()
    and is_sm100_supported()
    # EAGLE topk>1 spec runs the two-pass page-tree cascade, which the FA4
    # CUTLASS kernel aborts on at page_size>1. That path only works at
    # page_size==1, so skip the 128 auto-force for it and keep the default.
    and (view.speculative_eagle_topk or 0) <= 1
    # 以下行被删除:
    # # The full prefill CUDA graph runs the FA backend at page_size==1 only
    # # (#27988), so skip the 128 auto-force for it and keep the default.
    # and view.cuda_graph_config.prefill.backend != Backend.FULL
):
    logger.warning(
        f"FA4 backend only supports page size 128 for non-MLA model architectures, "
        f"changing page_size from {view.page_size} to 128."
    )
    return {"page_size": 128}
return {}

```

## 评论区精华

无 Review 讨论。PR body 明确指出该撤回是为了修复 CI 失败。

- 暂无高价值评论线程

## 风险与影响

- 风险：本 PR 是回退操作，风险较低。主要风险是丢失了 #30835 中旨在启用 Full Cuda Graph 与 page\_size=1 兼容的功能，但该功能本身导致 CI 失败，因此回退是必要的。
- 影响：对用户而言，FA4 后端在 full prefill CUDA graph 场景下恢复了 page\_size 自动设为 128 的行为，可能影响部分配置的 CUDA graph 启用条件。对系统稳定性有正面影响，CI 将恢复正常。
- 风险标记：回退原有功能

## 关联脉络

- PR #30835 [Tiny] Enable Full Cuda Graph with Page size = 1: 本 PR 回退了 #30835 的变更。