PR #26605 完整报告

sgl-project/sglang

[Log] include max_token_num and hidden_dim in FlashInfer workspace init log

合并时间：2026-06-02 04:26

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26605

执行摘要

- 一句话：日志增加 max_token_num 和 hidden_dim 信息
- 推荐动作：可快速合并，无需精读。

功能与动机

PR body 指出，服务器启动时输出的 "FlashInfer workspace initialized" 日志缺少工作区维度信息，而 max_token_num 和 hidden_dim 已在函数参数中，添加后有助于诊断 allreduce-fusion 缓冲区大小配置。

实现拆解

修改 python/sglang/srt/layers/flashinfer_comm_fusion.py 中 FlashInferWorkspaceManager.initialize 方法的日志输出行，在原有 rank、world_size、backend 之后追加 max_token_num 和 hidden_dim 的值。

关键文件：

- python/sglang/srt/layers/flashinfer_comm_fusion.py（模块 通信融合；类别 source；类型 core-logic）：日志行变更，追加工作区维度参数

关键符号：未识别

关键源码片段

python/sglang/srt/layers/flashinfer_comm_fusion.py

日志行变更，追加工作区维度参数

```
# 修改前：
# logger.info(
# f"FlashInfer workspace initialized for rank {rank}, "
# f"world_size {world_size}, backend {backend}"
# )

# 修改后：追加 max_token_num 和 hidden_dim，便于启动时诊断缓冲区尺寸
backend = getattr(self.workspace, "backend", "unknown")
logger.info(
    f"FlashInfer workspace initialized for rank {rank}, "
```

```
        f"world_size {world_size}, backend {backend}, "
        f"max_token_num {max_token_num}, hidden_dim {hidden_dim}"
    )
```

评论区精华

无实质讨论，reviewer 均直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：仅修改日志字符串，无任何功能逻辑变更，无回归、性能、安全或兼容性风险。
- 影响：影响极小，仅对运维人员在查看启动日志时提供更多诊断信息，用户和系统行为不受影响。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR