PR #26883 完整报告

sgl-project/sglang

[PP][Bugfix] Handle input_ids assignment in prepare_for_extend

合并时间：2026-06-01 14:43

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26883

执行摘要

- 一句话：修复 PP profiler 中 deferred H2D 后 input_ids 未赋值
- 推荐动作：值得合入，修复明确，逻辑可读且无副作用。建议验证 PP profiling 端到端测试通过。

功能与动机

PR body 指出这是一个“CPP bug (missing change of PR: #25945)”，即 #25945 重构了 prepare_for_extend 以支持 deferred H2D（将 input_ids 延迟到 GPU 上分配），但 PP profiler 路径（profile_and_init_predictor）未同步更新，导致 batch.input_ids 仍为 None 时被后续代码使用，引发错误。

实现拆解

1. 定位问题：在 python/sglang/srt/managers/scheduler_pp_mixin.py 的 profile_and_init_predictor 方法中，batch.prepare_for_extend() 调用后，由于 #25945 的变更，batch.input_ids 可能为 None（数据尚在 CPU 上）。
2. 添加回退逻辑：在 batch.prepare_for_extend() 之后、ForwardBatch.init_new 之前，插入一个条件判断：如果 batch.input_ids 为 None 且 batch.prefill_input_ids_cpu 不为 None，则将 prefill_input_ids_cpu 异步拷贝到 GPU 设备（non_blocking=True），并赋值给 batch.input_ids，同时将 batch.prefill_input_ids_cpu 置为 None 以释放内存。
3. 影响范围：仅影响 PP（Pipeline Parallelism）profiling 路径，不改变常规推理流程。

关键文件：

- python/sglang/srt/managers/scheduler_pp_mixin.py（模块 调度器；类别 source；类型 core-logic）：唯一修改的文件，在 profile_and_init_predictor 方法中增加了 deferred H2D 回退逻辑，修复了关键路径上的数据竞争问题。

关键符号：未识别

关键源码片段

python/sglang/srt/managers/scheduler_pp_mixin.py

唯一修改的文件，在 profile_and_init_predictor 方法中增加了 deferred H2D 回退逻辑，修复了关键路径上的数据竞争问题。

```
# 位于 profile_and_init_predictor 方法中
batch.prepare_for_extend()

# Resolve deferred H2D: prepare_for_extend 现在将 input_ids 延迟到 GPU
# 因此在之后需要检查 input_ids 是否为 None，并从 CPU 拷贝
if batch.input_ids is None and batch.prefill_input_ids_cpu is not None:
    # 异步拷贝到 GPU，避免阻塞
    batch.input_ids = batch.prefill_input_ids_cpu.to(
        self.device, non_blocking=True
    )
    batch.prefill_input_ids_cpu = None # 释放 CPU 内存
```

评论区精华

无 reviewer 评论，作者自行合并。CI 中仅作者触发 rerun 了相关 disaggregation PP 测试并通过。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低，修改仅影响 PP profiler 路径，且逻辑简单（从 CPU 拷贝到 GPU）。但需确认 prefill_input_ids_cpu 在 profiler 上下文中确实一直有效；若该属性未正确设置，则条件不会触发，但也不会引入新 crash。
- 影响：
  - 用户 / 系统：修复了启用 PP dynamic chunk profiling 时可能遇到的崩溃或错误，确保 PP profiler 正常工作。
  - 影响范围：仅影响使用 PP 且启用 profiling（profile_and_init_predictor）的用户，常规推理无影响。
  - 影响程度：中等，因为修复的是一个功能性回归。
  - 风险标记：PP 专属路径

关联脉络

- PR #25945 （原名未知）：引入 deferred H2D 重构，导致本 PR 修复的回归。