PR #32180 完整报告

sgl-project/sglang

[BugFix] Prevent TBO crash when return_logprob is enabled

合并时间：2026-07-26 15:07

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/32180

执行摘要

- 一句话：修复 TBO 开启 return_logprob 时的崩溃
- 推荐动作：该 PR 变更极小但关键，属于典型边界条件 bugfix，值得快速合入。建议后续为该逻辑添加单元测试，防止类似遗漏。

功能与动机

PR 描述指出，使用 return_logprob: true 参数发送单条请求时，TBO 服务器会崩溃。崩溃的根本原因是 ForwardBatch.filter_batch 方法会检查每个字段，若某个字段不为 None 但不在 output_dict 中，则抛出异常。当 return_logprob 启用时，父批次的 extend_input_logprob_token_ids_gpu 字段被赋值，但在将父批次拆分为多个 TBO 子批次时，该字段未被重置，导致子批次触发校验失败。

实现拆解

1. 定位问题：在 python/sglang/srt/batch_overlap/two_batch_overlap.py 的 filter_batch 方法中，构造子批次 ForwardBatch 的 output_dict 时，缺少 extend_input_logprob_token_ids_gpu 字段。由于父批次中该字段有值，子批次继承后未被覆盖，触发第 812-815 行的校验异常。
2. 修复：在 output_dict 的 dict 初始化中，在 token_ids_logprobs=None 之后新增一行 extend_input_logprob_token_ids_gpu=None，显式将子批次的该字段置为 None。
3. 影响：该修复仅一行代码，确保 TBO 拆分时清理此字段，避免校验失败。评审者指出此变更同时影响 NVIDIA 和 AMD 平台。

关键文件：

- python/sglang/srt/batch_overlap/two_batch_overlap.py（模块 调度器；类别 source；类型 core-logic）：修复所在文件，仅一行代码变更：在构造子批次 output_dict 时显式将 extend_input_logprob_token_ids_gpu 设为 None。

关键符号：filter_batch

关键源码片段

python/sglang/srt/batch_overlap/two_batch_overlap.py

修复所在文件，仅一行代码变更：在构造子批次 output_dict 时显式将 extend_input_logprob_token_ids_gpu 设为 None。

```
# python/sglang/srt/batch_overlap/two_batch_overlap.py
# 在 filter_batch 方法中构造子批次 ForwardBatch 的 output_dict 时，
# 显式将 extend_input_logprob_token_ids_gpu 设为 None，
# 避免因父批次中该字段有值而引发 filter_batch 校验异常（TBO + return_logprob 崩溃）。
output_dict.update(
    dict(
        # ... 其他字段 ...
        top_logprobs_nums=None,
        token_ids_logprobs=None,
        extend_input_logprob_token_ids_gpu=None, # <-- 新增：修复 TBO + return_logprob 崩溃
        next_token_logits_buffer=None,
        return_hidden_states_before_norm=False,
        # ... 其余字段 ...
    )
)
```

评论区精华

评审者 1am9trash 确认了问题机制：filter_batch 会拒绝任何未处理的非空字段，因此需要显式传递 next_token_logits_buffer=None（实际是 extend_input_logprob_token_ids_gpu=None）。该评审者还指出影响同时覆盖 NVIDIA 和 AMD 两个平台。另一位评审者 kkHuang-amd 也给予了 LGTM。讨论没有争议。

- 暂无高价值评论线程

风险与影响

- 风险：本次变更仅一行代码，将特定字段显式置为 None，风险极低。但回归风险在于：如果未来有逻辑依赖子批次的 extend_input_logprob_token_ids_gpu 继承父批次值，此修改可能破坏该逻辑。当前代码注释（TBO children start unplanned）表明子批次的该字段应当为空，因此风险可控。另外，没有添加单元测试覆盖该场景，建议补充。
- 影响：修复了 TBO 模式下 return_logprob 启用时的服务崩溃问题，使该功能组合恢复正常。影响范围仅限于使用 TBO 和 return_logprob 的用户场景，性能测试表明对 TTT 和 TTFT 几乎无影响。
- 风险标记：缺少测试覆盖

关联脉络

- 暂无明显关联 PR