PR #21260 完整报告

sgl-project/sglang

Add adjusted_filter_batch

合并时间：2026-03-26 10:59

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21260

执行摘要

本 PR 在 SamplingBatchInfo 类中添加了 adjusted_filter_batch 占位符方法，扩展了采样批处理的自定义能力，影响范围限于该类的子类。

功能与动机

PR body 未提供具体动机，但从代码变更推断，旨在为 SamplingBatchInfo 类提供一个扩展点，允许子类在过滤批次时执行自定义逻辑，增强灵活性。

实现拆解

修改了 python/sglang/srt/sampling/sampling_batch_info.py 文件：

- 新增 adjusted_filter_batch 方法，参数为 keep_indices 和 keep_indices_device，方法体为 pass，作为占位符。
- 在现有的 filter_batch 方法中添加调用 self.adjusted_filter_batch(keep_indices, keep_indices_device)，确保在过滤时触发。

评论区精华

review 过程中没有实质性讨论，仅由 ispobock 批准合并。

风险与影响

风险较低：新增方法是占位符，无实际实现，但子类若未正确覆盖可能导致行为不一致。当前无测试覆盖。 影响范围小：仅影响 SamplingBatchInfo 类及其子类，为开发人员提供自定义过滤逻辑的接口。

关联脉络

从历史 PR 分析中未发现直接相关 PR，这可能是一个独立的接口扩展，未来可能与其他采样或批处理优化 PR 关联。