PR #25819 完整报告

sgl-project/sglang

disagg prebuilt: drop dead prepare_for_extend shift

合并时间：2026-05-20 19:39

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25819

执行摘要

- 一句话：移除 disagg decode 中死代码，提取 EAGLE 预填充旋转函数
- 推荐动作：重构方向正确，代码风格（向量化、type hints）值得参考。建议精读 apply_eagle_prefill_input_rotation 的实现和解耦思路。社区 reviewer 可关注此类死代码清理。

功能与动机

PR body 说明：Drops the prepare_for_extend in-place shift on batch.input_ids in disagg decode prebuilt — no forward in _run_batch_prebuilt consumes it, and the next decode iter's verify overwrites batch.input_ids before any kernel reads it. Set batch.input_ids to the [bs] last-token tensor instead for cleaner filter/merge semantics.

实现拆解

1. 在 eagle_utils.py 中新增共享函数 apply_eagle_prefill_input_rotation，接收 ScheduleBatch 和 next_token_ids，执行向量化的左移旋转：rotated[:-1] = batch.input_ids[1:]，然后在每个请求的段尾 scatter 新 token。
2. 从 EagleDraftInput 类中删除 prepare_for_extend 方法（16 行循环实现），相关功能完全由新函数替代。
3. 更新调用方：在 eagle_worker.py 和 multi_layer_eagle_worker.py 的 forward_draft_extend 中，将 batch.spec_info.prepare_for_extend(batch) 替换为直接调用 apply_eagle_prefill_input_rotation(batch, next_token_ids)。
4. 在 disagg decode prebuilt 的 mixin 文件 decode_schedule_batch_mixin.py 中，移除对 prepare_for_extend 的调用及相关注释，因为该路径不消费移位后的 input_ids。
5. 测试 / 配置配套改动：无直接测试变更，但所有调用点均已覆盖。

关键文件：

- python/sglang/srt/speculative/eagle_utils.py（模块 EAGLE 工具；类别 source；类型 core-logic；符号 apply_eagle_prefill_input_rotation）：核心文件：新增向量化旋转函数，统一所有调用路径
- python/sglang/srt/speculative/eagle_info.py（模块 EAGLE 输入；类别 source；类型 core-logic；符号 prepare_for_extend）：删除旧的 prepare_for_extend 方法，移除死代码

- python/sglang/srt/disaggregation/decode_schedule_batch_mixin.py（模块 分离调度；类别 source；类型 core-logic）：移除无用的移位调用，简化逻辑
- python/sglang/srt/speculative/eagle_worker.py（模块 EAGLE 工作线程；类别 source；类型 core-logic）：替换为共享函数调用
- python/sglang/srt/speculative/multi_layer_eagle_worker.py（模块 多层 EAGLE 工作线程；类别 source；类型 core-logic）：替换为共享函数调用

关键符号：apply_eagle_prefill_input_rotation

评论区精华

PR 无外部 review comment，但 commit 历史显示作者多次迭代：包括类型注解调整、dtype 修复、注释精简等，反映出对代码清晰度和正确性的内部把关。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。移除了无用的移位操作，新函数语义与旧代码等价（向量化实现）。唯一注意点是 dtype 不一致：commit f26dae58 修复了 rotated[seg_ends] = next_token_ids.to(batch.input_ids.dtype)。如果其他依赖旧 prepare_for_extend 行为的路径未更新，可能出错，但已确认所有调用点均被替换。无性能退化风险，向量化可能稍优。
- 影响：影响范围仅限于 EAGLE 推测解码的 disagg decode 路径和 draft extend 路径。对外部用户行为无影响，内部逻辑更清晰。因使用 __future__annotations 和 TYPE_CHECKING，对导入性能略有优化。
- 风险标记：低风险，死代码移除，向量化重构

关联脉络

- PR #25774 disagg prebuilt related: Stacked on #25774，PR 基于该 disagg 基座构建