

PR #30461 完整报告

sgl-project/sglang

[DSV4] Fix draft SWA transfer for disaggregated MTP

合并时间: 2026-07-09 13:18

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30461>

执行摘要

- 一句话: 修复 DSV4 分离解码 draft SWA 状态传输缺失
- 推荐动作: 值得阅读以了解分离解码状态传输的组件扩展方式。设计上选择复用现有 SWA 传输类型而非新增类型, 保持了 forward 兼容性。测试模式值得参考。

功能与动机

当 prefill 和 decode 都启用 EAGLE 时, prefill 为 draft NextN 层构建 prompt SWA 状态, 但 draft 的 SWA 缓冲区未在分离传输协议中注册, 导致 decode 开始时缺少或 stale 的 draft prompt 状态, 直到新生成的 token 重新填充滑动窗口。(PR body 原话)

实现拆解

1. 导入依赖: 在 `python/sglang/srt/disaggregation/utils.py` 中增加对 `DeepSeekV4TokenToKVPool` 的导入, 用于后续类型检查。
2. 核心校验与注册: 在 `setup_state_kv_args` 函数末尾, 当目标池为 `DeepSeekV4TokenToKVPool` 且非 NPU 平台且 draft 池同为 `DeepSeekV4TokenToKVPool` 时, 执行:
 - 校验 draft 层为纯 SWA (`compression_ratios` 全为零), 否则抛出 `RuntimeError`。
 - 校验 target 和 draft 的 `_unified_kv` 标志一致。
 - 根据 unified 模式选择: 若 unified, 校验 SWA ring 几何 (`unified_swa_window`, `unified_swa_ring_size`, `unified_swa_pages`) 一致, 并从 draft 获取 `get_unified_swa_ring_buf_infos`; 若 paged, 校验 `full_to_swa_index_mapping` 共享, 并获取 `get_state_buf_infos`。
 - 使用 `append_state_component` 将 draft 缓冲区注册为额外的 `StateType.SWA_RING` (unified) 或 `StateType.SWA` (paged) 状态组件。
3. 测试覆盖: 在 `test/registered/unit/disaggregation/test_disaggregation_wire.py` 中新增 `TestDSV4DraftStateRegistration` 类, 利用模拟的 target 和 draft pool 对象验证 paged 和 unified 模式下 draft 状态组件正确添加且与 target 状态组件不混淆。

关键文件:

- `test/registered/unit/disaggregation/test_disaggregation_wire.py` (模块 分离测试; 类别 test; 类型 test-coverage; 符号 `_buf_infos`, `_make_dsv4_target`, `_make_dsv4_draft`,

TestDSV4DraftStateRegistration)：新增 draft 状态注册单元测试，覆盖 paged 和 unified 两种池模式，验证 draft 状态作为独立组件注册且与 target 状态不混淆。

- python/sglang/srt/disaggregation/utils.py (模块 分离工具；类别 source；类型 core-logic；符号 setup_state_kv_args)：核心变更文件，在 setup_state_kv_args 中添加 DSV4 draft 状态注册分支，包含校验和注册逻辑。

关键符号：setup_state_kv_args

评论区精华

评论者 ShangmingCai 指出，NPU 使用 NPUMLATokenToKVPool，因此显式的 is_npu() 检查可能冗余，但当前实现保留此检查以明确 NPU 路径未受影响。作者未回复，reviewer 随后批准合并，表示接受当前设计，后续 NPU 需要时可清理。

- is_npu 检查是否冗余 (design)：保留 is_npu 检查，后续 NPU 需要时可移除。

风险与影响

- 风险：主要风险在于非 DSV4 或非分离场景不受影响，但 draft 池注册失败时会抛出 RuntimeError，可能导致服务启动中断。然而这些校验在启动时触发，不会隐藏错误。另外，NPU 路径未修改，若 NPU 也需要此功能需单独补丁。测试使用模拟 pool，未覆盖所有真实池属性，但核心逻辑已验证。
- 影响：直接影响使用 DSV4 模型并启用 EAGLE 分离解码的用户：修复后 decode 阶段 draft 提案质量提升（早期接受长度提高 15-17%，全请求平均 AL 提高约 4%）。无功能退化，对其他模型和场景无影响。
- 风险标记：核心分离路径变更，缺少 NPU 覆盖，启动时校验可能中断服务

关联脉络

- PR #30408 Fix DSV4 HiSparse SWA tail allocation forwarding: 同为 DSV4 分离状态修复，修改了 disaggregation 模块和记忆池。
- PR #30636 [UnifiedTree]: Sync Replay SSM: 修复另一类状态传输（Mamba SSM）在分离中的缓存长度计算，与此 PR 关注点相似。