PR #28520 完整报告

sgl-project/sglang

[AMD] Fix deepseek-v4 mtp accept length issue

合并时间：2026-06-18 02:13

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28520

执行摘要

- 一句话：修复 DeepSeek-V4 MTP 接受长度下降至 ~2.17 的问题
- 推荐动作：建议阅读 get_unified_swa_loc 的修改（约 15 分钟），理解如何在保留缓存性能的同时检测变长多步场景。集成测试和 CI 配置可作为类似问题的测试范例。

功能与动机

PR 描述指出 unified_kv_triton 后端的平均接受长度仅 2.17，而 triton 后端为 3.04，导致 decode 吞吐严重受损。定位为 SWA 环写入位置缓存未考虑多步 draft decode，导致每步写入同一环槽，破坏 draft chain（token 2 后接受率骤降）。

实现拆解

1. 问题定位与修改：在 python/sglang/srt/layers/attention/deepseek_v4_backend_hip_radix.py 的 get_unified_swa_loc 方法中，添加 is_multistep_draft_decode 标志，当处于 decode 模式且 speculative_num_steps > 1 时，绕过缓存，从实时 positions 重新计算环位置。
2. 保持快速路径：对非多步 draft 路径（prefill、extend、单步 decode）仍使用缓存，避免计算开销。
3. 新增集成测试：创建 test/registered/amd/test_deepseek_v4_pro_fp4_mtp.py，包含 GSM8K 精度测试和 MTP 接受长度测试，在 8-GPU MI35x 上验证两种后端。
4. CI 配置：在 .github/workflows/nightly-test-amd-rocm720.yml 添加 nightly-8-gpu-mi35x-deepseek-v4-pro-mtp-rocm720 作业，分别以 unified_kv_triton 和 triton 后端运行测试。

关键文件：

- python/sglang/srt/layers/attention/deepseek_v4_backend_hip_radix.py（模块 注意力后端；类别 source；类型 core-logic；符号 get_unified_swa_loc）：核心修复文件，修改了 get_unified_swa_loc 方法，添加多步 draft decode 检测。
- test/registered/amd/test_deepseek_v4_pro_fp4_mtp.py（模块 集成测试；类别 test；类型 test-coverage；符号 TestDeepseekV4ProFp4MTP, setUpClass, tearDownClass, test_a_gsm8k）：新增集成测试，验证 DeepSeek-V4-Pro FP4 在两种注意力后端下的精度和接受长度。

- .github/workflows/nightly-test-amd-rocm720.yml（模块 CI 配置；类别 infra；类型 infrastructure）：添加 nightly CI 作业运行新测试，覆盖 unified_kv_triton 和 triton 后端。

关键符号：get_unified_swa_loc, TestDeepseekV4ProFp4MTP.setUpClass, TestDeepseekV4ProFp4MTP.test_a_gsm8k, TestDeepseekV4ProFp4MTP.test_b_bs_1_speed

关键源码片段

python/sglang/srt/layers/attention/deepseek_v4_backend_hip_radix.py

核心修复文件，修改了 get_unified_swa_loc 方法，添加多步 draft decode 检测。

```
def get_unified_swa_loc(self, forward_batch: ForwardBatch) -> torch.Tensor:
    """SWA ring write target for unified_kv, shared by all layers.
    ...
    """
    positions = forward_batch.positions
    core = getattr(self.forward_metadata, "core_attn_metadata", None)
    unified = getattr(core, "unified", None) if core is not None else None
    cached = unified.swa_loc if unified is not None else None
    # 新增检测：多步 draft decode 时绕过缓存
    is_multistep_draft_decode = (
        forward_batch.forward_mode.is_decode_or_idle()
        and self.speculative_num_steps > 1
    )
    if (
        cached is not None
        and not forward_batch.forward_mode.is_idle()
        and cached.shape[0] == positions.shape[0]
        and not is_multistep_draft_decode
    ):
        result = cached
    else:
        ring = self.token_to_kv_pool.unified_swa_ring_size
        req_slot = forward_batch.req_pool_indices.to(torch.int64)
        if req_slot.shape[0] != positions.shape[0]:
            req_slot = req_slot.repeat_interleave(
                positions.shape[0] // req_slot.shape[0]
            )
        result = (req_slot * ring + positions.to(torch.int64) % ring).to(torch.int32)
    return result
```

评论区精华

代码审查未引发技术讨论，维护者 HaiShaw 直接批准。Gemini Code Assist 的自动评论未提供实质性反馈。

- 暂无高价值评论线程

风险与影响

- 风险：主要风险在 is_multistep_draft_decode 检测条件若存在边界情况（如 idle 模式误判）可能导致性能回退，但不影响正确性。修改仅影响 unified_kv_triton 后端，对 triton 后端和其他模型无影响。测试覆盖 two 种后端，回归风险较低。
- 影响：对 AMD 平台 DeepSeek-V4 MTP 用户：接受长度从 2.17 提升至 3.08，总 token 吞吐从 6355.88 增至 7324.37 tok/s（+15%）。新增 CI 套件防止后续回归。团队获得一个清晰的缓存绕过模式，可供其他多步 draft 场景参考。
- 风险标记：AMD 专属变更，核心路径变更，已添加测试覆盖

关联脉络

- 暂无明显关联 PR