PR #36915 完整报告

sgl-project/sglang

[AMD] Fix eager metadata for AITER EAGLE draft extend

合并时间：2026-08-29 13:56

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/36915

执行摘要

- 一句话：修复 AITER EAGLE draft extend eager 元数据错误
- 推荐动作：该 PR 值得精读，因为它修复了特定硬件和模型组合下的关键稳定性问题，并展示了元数据传递的正确模式。值得关注的设计决策是使用 generate_attn_arg_prefill() 返回值构建 metadata。

功能与动机

PR body 指出问题复现于 AMD MI355X 上，Qwen3.5-397B-A17B-MXFP4，TP2，AITER attention，EAGLE 3 步 4 token，FP8 KV cache。日志显示 q_shape=(4, 4096) 但 cu_seqlens_q=[0, 80]，导致 unified_attention 处理 80 行 query 而实际只有 4 行，产生越界写入和内存访问故障，甚至出现 IndexError: list index out of range 和 HSA_STATUS_ERROR_MEMORY_APERTURE_VIOLATION 导致的 Python 崩溃。

实现拆解

1. 修改 python/sglang/srt/layers/attention/aiter_backend.py 中 init_forward_metadata 的非 MLA draft extend 分支，用 forward_batch.spec_info.generate_attn_arg_prefill() 返回的 token 级 kv_indices, kv_indptr, qo_indptr 填充 ForwardMetadata。
2. 将 ForwardMetadata 的 max_q_len 设为 spec_info.num_tokens_per_req，并保留计算的 max_kv_len。
3. 在 forward_extend 中，将 unified_attention 的 cu_seqlens_q 参数从 self.qo_indptr[: bs + 1] 改为 self.forward_metadata.qo_indptr，确保 query 序列边界与当前 query 张量一致。

关键文件：

- python/sglang/srt/layers/attention/aiter_backend.py（模块 注意力；类别 source；类型 core-logic；符号 init_forward_metadata, forward_extend）：核心修复文件，修改 eager draft extend 的 ForwardMetadata 构建和 unified_attention 调用。

关键符号：init_forward_metadata, forward_extend

关键源码片段

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

核心修复文件，修改 eager draft extend 的 ForwardMetadata 构建和 unified_attention 调用。

```
# 修复前使用 indices_updater_prefill，导致 qo_indptr 未设置
# 现在直接使用 spec_info.generate_attn_arg_prefill 的返回值

kv_indices, kv_indptr, qo_indptr, _ = (
    forward_batch.spec_info.generate_attn_arg_prefill(
        forward_batch.req_pool_indices,
        forward_batch.seq_lens,
        forward_batch.seq_lens_sum,
        self.req_to_token,
    )
)
self.forward_metadata = ForwardMetadata(
    kv_indptr,
    kv_indices,
    qo_indptr, # 以前传 None，现在传真正的 query indptr
    None,
    forward_batch.spec_info.num_tokens_per_req, # 正确的 max_q_len
    max_kv_len,
)

# forward_extend 中，将 cu_seqlens_q 改为传入 forward_metadata 的 qo_indptr
cu_seqlens_q=self.forward_metadata.qo_indptr,
```

评论区精华

reviewer yichiche 评论："Nice catch, we see this issue raised in agent mode testing, thanks for the quick fix. LGTM." 另一位 reviewer HaiShaw 也批准了该 PR。无其他讨论。

- 修复必要性 (question): 批准修复

风险与影响

- 风险：风险较低，仅修改非 MLA draft extend eager 路径，不影响 CUDA Graph 路径和 MLA 路径。但缺少单元测试，修复可能因未来重构而回归。建议在后续添加针对该路径的测试。
- 影响：影响范围限于使用 AITER unified-attention 的非 MLA EAGLE 模型在 eager 模式或超 batch 时的稳定性。修复后避免内存访问错误和调度器异常，提升 AMD 平台的可靠性。团队应关注相关测试覆盖的补充。
- 风险标记：缺少测试覆盖，特定硬件路径变更

关联脉络

- PR #30105 Enable AITER unified draft extend by default: 本 PR 修复了 #30105 引入的回归