PR #28026 完整报告

sgl-project/sglang

[Bugfix][Spec] Fix multi-layer EAGLE DRAFT_EXTEND_V2 attn-TP logprob metadata capture

合并时间：2026-06-13 10:26

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28026

执行摘要

- 一句话：修复多层 EAGLE attn-TP 的 logprob metadata 值错误
- 推荐动作：此 PR 值得精读，作为理解 speculative decoding 中 tensor parallelism 与 logprob metadata 一致性的典型案例。

功能与动机

修复 Issue #26416 报告的错误：MultiLayerEagleDraftExtendCudaGraphRunner.get_forward_batch() 在 require_attn_tp_gather 分支中设置 global_num_tokens_for_logprob_cpu = [bs]，但 DRAFT_EXTEND_V2 为所有 draft-extend tokens 计算 logits，正确的 metadata 应为 num_tokens = bs * num_tokens_per_bs。该错误与 global_num_tokens_cpu、global_dp_buffer_len、replay fill 以及 single-layer runner 的行为不一致。

实现拆解

1. 核心源码路径 - python/sglang/srt/speculative/multi_layer_eagle_draft_extend_cuda_graph_runner.py（core-logic）：源码主路径；包含 配置键调整；+2/-1

关键文件：

- python/sglang/srt/speculative/multi_layer_eagle_draft_extend_cuda_graph_runner.py（模块 推测解码；类别 source；类型 core-logic）：核心修复文件：在 get_forward_batch 的 require_attn_tp_gather 分支中将 global_num_tokens_for_logprob_cpu 从 [bs] 改为 [num_tokens]，共 2 行新增 1 行删除。

关键符号：未识别

关键源码片段

python/sglang/srt/speculative/multi_layer_eagle_draft_extend_cuda_graph_runner.py

核心修复文件：在 get_forward_batch 的 require_attn_tp_gather 分支中将 global_num_tokens_for_logprob_cpu 从 [bs] 改为 [num_tokens]，共 2 行新增 1 行删除。

```
def get_forward_batch(self, bs: int) -> ForwardBatch:
    buffers = self.buffers
    num_tokens = bs * self.num_tokens_per_bs # 总 token 数
```

```
# ... 省略其他切片操作 ...
if self.require_mlp_tp_gather:
    global_num_tokens_cpu = [num_tokens] * self.dp_size
    global_num_tokens_for_logprob_cpu = [num_tokens] * self.dp_size
elif self.require_attn_tp_gather:
    global_num_tokens_cpu = [num_tokens]
    # 修复：DRAFT_EXTEND_V2 对所有 token 产生 logits，应与 mlp 分支一致
    global_num_tokens_for_logprob_cpu = [num_tokens] # 原为 [bs]，导致 metadata 错误
else:
    global_num_tokens_cpu = None
# ... 后续复制到 GPU tensor 以供 Tensor Parallelism 使用 ...
```

评论区精华

无有效 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅修改一行赋值，且修复方向明确（将 [bs] 改为 [num_tokens]），与 mlp 分支和 replay fill 逻辑对齐。未引入新的依赖或控制流。
- 影响：影响范围限定在多 GPU attention-TP 场景下的 multi-layer EAGLE 推理，修复后发现 logprob 和 logit 值正确。对单 GPU 或 MLP-TP 路径无影响。
- 风险标记：核心路径变更，缺少测试覆盖

关联脉络

- PR #28093 [Spec] Move draft-extend prep to EagleDraftWorkerBase; unify prepare_for_* names: 同一功能线（speculative decoding）的近期重构，影响了相同的 multi_layer_eagle_worker 文件，有助于理解整体演进。
- PR #28081 [refactor] Fold FrozenKVMTPCudaGraphRunner onto the shared DecodeCudaGraphRunner base: 同为 speculative decoding 的图形运行器重构，与本 PR 关注的多层 EAGLE 图形运行器相关。
- PR #28096 [Spec] Fix EagleDraftWorker draft-extend attn backend assignment: 另一个 speculative decoding 的 attn 后端错误修复，与 logprob metadata 的上下文有关。