

PR #27647 完整报告

sgl-project/sglang

[sgl] Fix kimi-k2.5 EAGLE3 MLA draft embeds for batched MM prefill

合并时间: 2026-06-10 02:26

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27647>

执行摘要

- 一句话: 修复 kimi-k2.5 EAGLE3 MLA 批处理 draft 嵌入计算
- 推荐动作: 建议合入, 因为修复了一个静默的正确性错误且改动极小。开发团队可在后续 PR 中补充针对多模态批处理的单元测试。

功能与动机

PR 描述指出: 当前代码仅修补 flat extend buffer 的最后位置, 假设 batch size 为 1。对于更大的 batch, 请求 0..N-2 在其每个请求的最后一个 token 位置上静默保留了过时的 mm_input_embeds——不会崩溃, 但会导致错误的 draft 输入和降低的接受率。

实现拆解

1. 识别问题: 在 kimi_k25_eagle3.py 的 forward 函数中, 当处于多模态扩展模式且非 draft_extend 时, 原有逻辑通过 `torch.cat([embeds[:-1], self.embed_tokens(input_ids[-1].unsqueeze(0))])` 仅更新最后一个 token 的嵌入, 仅适用于 batch size 为 1 的情况。
2. 计算每个请求的 last-token 索引: 使用 `forward_batch.extend_start_loc + forward_batch.extend_seq_lens - 1` 计算出每个序列在 flat extend buffer 中的最后一个 token 位置向量 last_indices。
3. 原地赋值: 通过 `embeds[last_indices] = self.embed_tokens(input_ids[last_indices])` 一次性将正确的嵌入写入所有请求的最后一个 token 位置, 无需临时张量拼接, 同时避免 old 代码中的 `torch.cat` 带来的全张量分配。

关键文件:

- python/sglang/srt/models/kimi_k25_eagle3.py (模块 推测解码; 类别 source; 类型 data-contract; 符号 forward): 唯一修改的文件, 包含核心 forward 方法中 draft 嵌入的批处理修复。

关键符号: forward

关键源码片段

[python/sglang/srt/models/kimi_k25_eagle3.py](#)

唯一修改的文件, 包含核心 forward 方法中 draft 嵌入的批处理修复。

```
# 原代码 (base) 片段:
# embeds = torch.cat([embeds[:-1], self.embed_tokens(input_ids[-1].unsqueeze(0))])
# 新代码 (head) 片段:
last_indices = (
    forward_batch.extend_start_loc + forward_batch.extend_seq_lens - 1
).long()
embeds[last_indices] = self.embed_tokens(input_ids[last_indices])
```

评论区精华

PR 没有实质性的 reviewer 讨论。唯一的 comment 来自 gemini-code-assist[bot] 的自动总结，无异议。Qiaolin-Yu 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险低：只有 7 行改动 (+4/-3)，逻辑清晰，且通过索引计算避免了假设 batch=1 的问题。没有引入新的分支或配置。但未附带单元测试，回归风险依赖统一测试覆盖。
- 影响：影响范围：仅影响 kimi-k2.5 EAGLE3 MLA 模型在多模态批处理 (batch > 1) 扩展步骤时的 draft 嵌入，修复后 draft 输入正确，接受率恢复预期。对于单请求或非多模态场景无影响。性能上避免了 torch.cat 的全张量分配，可能略有提升。
- 风险标记：缺少测试覆盖

关联脉络

- PR #25980 Fix spec v2 stop output boundary: 同属 speculative decoding 路径的 bugfix 修复。
- PR #23802 fix: stop-string check misses early matches during speculative decoding: 同属 speculative decoding 路径的 bugfix 修复。