PR #23613 完整报告

sgl-project/sglang

[sgl] copy mm_input in piecewise cuda graph when eagle3 is on

合并时间：2026-04-28 04:35

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23613

执行摘要

- 一句话：修复 EAGLE3 分段 CUDA Graph 下 mm_input 丢失
- 推荐动作：值得精读：该 PR 修复了一个涉及分段 CUDA Graph 和推测解码交互的边界问题，代码简洁且有明确条件守卫，是理解 SGLang 推测解码与 CUDA Graph 如何协作的良好范例。

功能与动机

PR body 指出当分段 CUDA Graph 启用时，需要从 CUDA Graph 输出缓冲区拷贝 mm_input 以使 EAGLE3 正常工作。speculative draft 的预填充路径 (eagle_worker_v2._draft_extend_for_prefill) 读取此 LogitsProcessorOutput 中的 mm_input_embeds 以复用目标编码器的嵌入，而非重新编码多模态占位符 token ID。

实现拆解

修改 python/sglang/srt/model_executor/piecewise_cuda_graph_runner.py 中 replay 方法：

1. 在 if isinstance(output, LogitsProcessorOutput) 分支内新增代码：初始化 mm_input_embeds = None。
2. 检查是否为推测解码模式 (self.model_runner.spec_algorithm.is_speculative()) 且输出中包含 mm_input_embeds，若是则从中切片出前 raw_num_tokens 个元素赋值给 mm_input_embeds。
3. 构造 LogitsProcessorOutput 时，将 mm_input_embeds 作为参数传入。 该变更仅影响分段 CUDA Graph 模式下的推测解码目标模型输出构建，不影响其他分支。

关键文件：

- python/sglang/srt/model_executor/piecewise_cuda_graph_runner.py（模块 推理引擎；类别 source；类型 data-contract；符号 replay）：核心修改文件：在 replay 方法的 LogitsProcessorOutput 构造逻辑中添加了 mm_input_embeds 的拷贝，确保分段 CUDA Graph 下 EAGLE3 多模态嵌入正确传递。

关键符号：replay

关键源码片段

python/sglang/srt/model_executor/piecewise_cuda_graph_runner.py

核心修改文件：在 replay 方法的 LogitsProcessorOutput 构造逻辑中添加了 mm_input_embeds 的拷贝，确保分段 CUDA Graph 下 EAGLE3 多模态嵌入正确传递。

```
# python/sglang/srt/model_executor/piecewise_cuda_graph_runner.py
# 在 replay 方法中，处理 LogitsProcessorOutput 时增加 mm_input_embeds 拷贝逻辑
if isinstance(output, LogitsProcessorOutput):
    # 保留 mm_input_embeds 用于推测解码：
    # speculative draft 的预填充路径 (eagle_worker_v2._draft_extend_for_prefill)
    # 从 LogitsProcessorOutput 中读取 mm_input_embeds，复用目标编码器嵌入，
    # 避免重新编码多模态占位符 token ID。当分段 CUDA Graph 启用时，
    # 需要从 CUDA Graph 输出缓冲区拷贝此字段。
    mm_input_embeds = None
    if (
        self.model_runner.spec_algorithm.is_speculative() # 仅在推测解码模式下
        and output.mm_input_embeds is not None # 且输出中存在 mm_input_embeds
    ):
        # 只保留原始 token 数量对应的嵌入，去除 CUDA Graph 填充部分
        mm_input_embeds = output.mm_input_embeds[: self.raw_num_tokens]
    return LogitsProcessorOutput(
        next_token_logits=output.next_token_logits[: self.raw_num_tokens],
        hidden_states=(
            output.hidden_states[: self.raw_num_tokens]
            if output.hidden_states is not None
            else None
        ),
        mm_input_embeds=mm_input_embeds, # 新增字段，传递切片后的嵌入
    )
```

评论区精华

本 PR 无显式 review 讨论或评论，但有 reviewer Qiaolin-Yu 的 APPROVAL。CI 经过两次 /rerun-failed-ci 后通过。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低：变更仅在分段 CUDA Graph 模式且推测解码算法（如 EAGLE3）启用时生效，通过 self.model_runner.spec_algorithm.is_speculative() 条件限制；未影响其他代码路径。但缺少针对该场景的单元测试覆盖，回归依赖集成测试。
- 影响：影响范围小，仅适用于 EAGLE3 结合分段 CUDA Graph 的场景。修复了多模态输入在推测解码下精度可能下降的问题，提升了该路径的正确性。不影响非推测解码、非分段 CUDA Graph 或非多模态模型。
- 风险标记：缺少测试覆盖

关联脉络

- PR #22997 [Whisper] Automatic language detection via structured generation: 同样涉及多模态输入和推测解码路径，与本 PR 有功能关联。