PR #25711 完整报告

sgl-project/sglang

Expose can-run-cuda-graph as a regular property on the embedding result

合并时间：2026-05-19 09:18

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25711

执行摘要

- 一句话：EmbeddingBatchResult 显式暴露 can_run_cuda_graph 属性
- 推荐动作：可快速合并，作为重构链中的一环，为未来统一类型处理打下基础。

功能与动机

EmbeddingBatchResult 没有声明 can_run_cuda_graph 字段，导致 process_batch_result_prefill 等函数需要通过 getattr(result, "can_run_cuda_graph", False) 来读取，这种隐式回退对类型系统不友好。由于 Embedding 前向从不运行 CUDA Graph，其值恒为 False，应通过显式属性暴露该事实。

实现拆解

1. python/sglang/srt/managers/utils.py: 在 EmbeddingBatchResult 数据类中添加 can_run_cuda_graph 属性，使用 @property 装饰器，始终返回 False。
2. python/sglang/srt/dllm/mixin/scheduler.py: 将 getattr(result, "can_run_cuda_graph", False) 替换为 result.can_run_cuda_graph 直接访问。
3. python/sglang/srt/managers/scheduler_components/batch_result_processor.py: 同样将 getattr 调用替换为直接属性访问。
4. python/sglang/srt/disaggregation/prefill.py: 同样将 getattr 调用替换为直接属性访问。

所有改动均为机械性替换，不引入新依赖或运行时状态。

关键文件：

- python/sglang/srt/managers/utils.py（模块 结果处理器；类别 source；类型 core-logic；符号 can_run_cuda_graph）：在 EmbeddingBatchResult 数据类中新增 can_run_cuda_graph property，是核心变更点。
- python/sglang/srt/dllm/mixin/scheduler.py（模块 调度器；类别 source；类型 core-logic）：DLLM 调度器对 prefill batch 结果的处理入口，统一了属性访问方式。
- python/sglang/srt/managers/scheduler_components/batch_result_processor.py（模块 结果处理器；类别 source；类型 core-logic）：Batch 结果处理器，处理 prefill 结果后上报指标，同样统一了属性访问。
- python/sglang/srt/disaggregation/prefill.py（模块 拆分配置；类别 source；类型 core-logic）：分离式 prefill 处理路径，同样统一了属性访问方式。

关键符号：EmbeddingBatchResult.can_run_cuda_graph

关键源码片段

python/sglang/srt/managers/utils.py

在 EmbeddingBatchResult 数据类中新增 can_run_cuda_graph property，是核心变更点。

```
@dataclass
class EmbeddingBatchResult:
    """Embeding / classification forward 的结果。"""
    embeddings: torch.Tensor
    pooled_hidden_states: Optional[torch.Tensor] = None
    copy_done: Optional[torch.cuda.Event] = None

    @property
    def can_run_cuda_graph(self) -> bool:
        # Embedding 模型从不运行 CUDA Graph，因此始终返回 False
        return False

    def copy_to_cpu(self):
        # ...
```

评论区精华

该 PR 仅有 1 条来自 gemini-code-assist 的评论，提示已到达每日配额限制，未讨论技术细节。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：改动为纯机械替换，逻辑等价。所有调用处原来的 getattr(result, "can_run_cuda_graph", False) 在 EmbeddingBatchResult 实例上始终返回 False，现在通过 property 显式返回 False，行为无变化。但需确保所有 EmbeddingBatchResult 类型的结果变量都被正确捕获，不存在残留的 getattr 调用。
- 影响：影响范围小：仅涉及 Embedding 相关的 prefill 结果处理路径，对 GenerationBatchResult 无影响。代码可读性和类型检查友好度提升。
- 风险标记：缺少测试覆盖

关联脉络

- PR #25724 Return a mamba tracking entry from the cache lookup instead of mutating caller lists: 同为重构系列，对结果类型进行封装，提升类型安全性。
- PR #25712 Pack scattered request logprob state into a dedicated container: 类似地将分散状态封装为专用容器，属于同一重构链条。