PR #26779 完整报告

sgl-project/sglang

[core] Compute dimensions/return_pooled_hidden_states in ForwardBatch.init_new

合并时间：2026-05-31 08:38

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26779

执行摘要

- 一句话：将 dimensions 等计算移至 ForwardBatch.init_new
- 推荐动作：值得阅读该 PR 以了解如何将只被下游消费者使用的计算从调度器迁移至前向初始化阶段，这是一种典型的分层清理方法。

功能与动机

根据 PR 描述，dimensions 和 return_pooled_hidden_states 两个字段只在 ForwardBatch 中被读取，调度器内部（ScheduleBatch 或 downstream 模块）并不使用。因此将它们的计算从 prepare_for_extend 移到 ForwardBatch.init_new，直接基于 batch.reqs 计算，从而删除 ScheduleBatch 上的冗余状态。

实现拆解

1. 在 forward_batch_info.py 中新增 _maybe_init_prefill_only 方法，从 batch.reqs 聚合 dimensions、return_pooled_hidden_states 和 multi_item_delimiter_indices；通过 is_prefill_only 门控仅对 prefill-only 批次设置。
2. 在 ForwardBatchInfo.init_new 中，在构造实例后调用 ret._maybe_init_prefill_only(batch)，替换原来直接从 batch 复制值的代码。
3. 在 schedule_batch.py 中删除 dimensions、return_pooled_hidden_states、multi_item_delimiter_indices 字段定义，并从 prepare_for_extend 方法中移除对应的计算逻辑。
4. 调整参数传递：从 init_new 的构造函数调用中移除 dimensions 和 return_pooled_hidden_states，改为通过 helper 方法注入。

关键文件：

- python/sglang/srt/model_executor/forward_batch_info.py（模块 前向批处理；类别 source；类型 data-contract；符号 _maybe_init_prefill_only）：新增 _maybe_init_prefill_only 方法并在 init_new 中调用，核心重构发生于此。
- python/sglang/srt/managers/schedule_batch.py（模块 调度器；类别 source；类型 core-logic；符号 prepare_for_extend）：删除三个字段和对应的计算逻辑，简化调度器。

关键符号：ForwardBatchInfo.init_new, ForwardBatchInfo._maybe_init_prefill_only, ScheduleBatch.prepare_for_extend

关键源码片段

python/sglang/srt/model_executor/forward_batch_info.py

新增 _maybe_init_prefill_only 方法并在 init_new 中调用，核心重构发生于此。

```
def _maybe_init_prefill_only(self, batch):
    """从 batch.reqs 派生每个请求的字段，用于非生成（embedding/reward/
    scoring）前向传播；否则为空操作。"""
    # 仅在 prefill-only 批次中设置
    if not self.is_prefill_only:
        return

    # 只有 Matryoshka 模型且至少有一个请求指定了 dimensions 时，才构建 dimensions 列表
    if batch.model_config.is_matryoshka and any(
        r.dimensions is not None for r in batch.reqs
    ):
        self.dimensions = [
            # 若请求未指定则使用模型的 hidden_size 作为默认值
            r.dimensions if r.dimensions else batch.model_config.hidden_size
            for r in batch.reqs
        ]

    # 对 pooled hidden states，只要任意一个请求要求就返回
    self.return_pooled_hidden_states = any(
        r.return_pooled_hidden_states for r in batch.reqs
    )

    # 多项目评分（MIS）所需的 delimiter indices，仅在 --enable-mis 启用时才计算
    if get_global_server_args().enable_mis and any(
        r.multi_item_delimiter_indices is not None for r in batch.reqs
    ):
        # score 端点保证所有请求都携带 delimiter indices
        assert all(
            r.multi_item_delimiter_indices is not None for r in batch.reqs
        ), "MIS batch 必须每个请求都有 delimiter indices"
        self.multi_item_delimiter_indices = [
            torch.tensor(r.multi_item_delimiter_indices, dtype=torch.int64)
            for r in batch.reqs
        ]

# 在 ForwardBatchInfo.init_new 尾部的调用：
ret = ForwardBatchInfo(...)
ret._maybe_init_prefill_only(batch)
```

评论区精华

无实质讨论，审阅者 gemini-code-assist[bot] 评论表示变更干净正确，没有进一步建议。

- Code Review by gemini-code-assist (other): No further feedback.

风险与影响

- 风险：主要风险是 multi_item_delimiter_indices 原本在 prepare_for_extend 中无条件生成，现在仅在 is_prefill_only 为 True 时生成。如果 decode 路径中需要此字段，将导致 None 访问。但根据上下文，该字段仅被 ForwardBatch 在 prefill-only 时使用，因此行为保持。另外，is_prefill_only 门控条件可能与上游调度器行为耦合，需确保调度器在所有 prefill 场景下正确设置此标志。缺少针对迁移后逻辑的测试覆盖。
- 影响：影响范围小，仅涉及内部两个源文件。外部 API 和用户行为无变化。团队内代码维护性提升，ScheduleBatch 状态减少。
- 风险标记：核心路径变更，缺少测试覆盖，条件分支变更

关联脉络

- PR #25945 [Scheduler] Defer prefill input_ids H2D to forward stream, unify resolve via future_map: 同样是调度器与前向阶段职责划分的重构，将 H2D 复制推迟到 forward 流，统一 resolver 路径，体现了相似的演进方向。