PR #26097 完整报告

sgl-project/sglang

[VLM] try to reuse precomputed padded input ids in scheduler instead of padding

合并时间：2026-05-25 10:27

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26097

执行摘要

- 一句话：调度器复用预计算的 padd 输入 ids，避免重复 padd 计算
- 推荐动作：值得精读，展示了如何通过“尝试 - 回退”模式在现有流程中插入预计算优化，设计简洁且侵入性低。团队成员可关注 _try_apply_padded_mm_input_ids 的边界条件处理及后续是否需补充测试。

功能与动机

PR body 说明：旨在调度器中复用已预计算的 MultimodalInputs.padded_input_ids，避免每次重复执行模型特定的 padd 工作，同时保留 session 前缀。

实现拆解

1. 在 scheduler.py 的 Scheduler 类中新增静态方法 _try_apply_padded_mm_input_ids(recv_req, req, image_inputs) -> bool。
2. 该方法检查 image_inputs.padded_input_ids 是否存在且长度与 recv_req.input_ids 一致，若一致则直接拼接到 req.origin_input_ids（考虑 session 前缀），返回 True；否则返回 False，触发回退逻辑。
3. 在 handle_generate_request 和 handle_embedding_request 多模态处理分支中，将原本的 if self.pad_input_ids_func: 改为 if not self._try_apply_padded_mm_input_ids(...) and self.pad_input_ids_func:，实现“优先复用，失败后 padd”的语义。

关键文件：

- python/sglang/srt/managers/scheduler.py（模块 调度器；类别 source；类型 core-logic；符号 _try_apply_padded_mm_input_ids）：核心变更文件，新增 _try_apply_padded_mm_input_ids 方法并修改 handle_generate_request 和 handle_embedding_request 中的 padd 逻辑。

关键符号：_try_apply_padded_mm_input_ids

关键源码片段

python/sglang/srt/managers/scheduler.py

核心变更文件，新增 _try_apply_padded_mm_input_ids 方法并修改 handle_generate_request 和 handle_embedding_request 中的 padd 逻辑。

```
# python/sglang/srt/managers/scheduler.py
@staticmethod
def _try_apply_padded_mm_input_ids(recv_req, req, image_inputs) -> bool:
    """
    尝试复用 MultimodalInputs.padded_input_ids 到 req.origin_input_ids，
    避免每次调用模型特定的 pad_input_ids_func 重新 padd。
    返回 True 表示成功复用，False 表示需要回退到原有 padd 逻辑。
    """
    padded_input_ids = image_inputs.padded_input_ids
    # 若预计算数据缺失或接收的 input_ids 为空，无法复用
    if padded_input_ids is None or recv_req.input_ids is None:
        return False

    recv_input_len = len(recv_req.input_ids)
    # 预计算长度必须与接收的 input_ids 长度一致
    if len(padded_input_ids) != recv_input_len:
        return False

    # 计算 session 前缀长度（origin_input_ids 可能包含之前 session 的 token）
    prefix_len = len(req.origin_input_ids) - recv_input_len
    if prefix_len < 0:
        return False

    padded_input_ids = array("q", padded_input_ids)
    if prefix_len == 0:
        req.origin_input_ids = padded_input_ids
    else:
        # 保留前缀，拼接新的 padded ids
        req.origin_input_ids = req.origin_input_ids[:prefix_len] + padded_input_ids
    return True
```

评论区精华

Gemini Code Assist bot 指出了两个高优先级回归：在 python/sglang/srt/models/qwen3_vl.py 的 get_image_feature 和 get_video_feature 中，pixel_values 的显式类型转换 .type(self.visual.dtype) 被移除，可能导致 dtype 不匹配错误。它建议恢复为 .to(self.visual.dtype)。不过本次 PR 的 diff 不涉及该文件，可能是来自其他变更的误报。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更集中在 scheduler.py 的两个调用点，逻辑是“先尝试复用，失败则回退到原有路径”，因此不会破坏现有行为。潜在风险：如果 padded_input_ids 内容与 pad_input_ids_func 输出存在语义差异（比如 session 场景下前缀计算有误），可能导致错误，但长度校验提供了一定保护。缺少测试覆盖。
- 影响：影响范围：仅影响多模态请求处理路径，包括 generate 和 embedding 请求。性能影响：在多模态输入已预计算 padded ids 的场景下可减少 padd 计算开销。对用户、系统、

团队影响较小，属于渐进式优化。

- 风险标记：缺少测试覆盖

关联脉络

- PR #26101 [VLM] accept precomputed multimodal metadata: 该 PR 在 producer 侧启用预计算 multimodal metadata，本 PR 则是在 scheduler（consumer）侧消费该预计算的 padded_input_ids，两者形成上下游。