PR #21269 完整报告

sgl-project/sglang

Fix sessions with mm inputs

合并时间：2026-03-27 08:38

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21269

执行摘要

- 一句话：修复会话中多模态输入的内存清除和偏移调整，恢复视觉会话测试。
- 推荐动作：建议工程师精读 session_controller.py 中的偏移调整逻辑和测试文件的变更，关注多模态输入在会话中的生命周期管理。

功能与动机

根据 PR body，作者指出：'We can only clear multimodal inputs from the request if it's not part of any session. Otherwise, it may be reused later. - We need to adjust mm offsets if bos token is trimmed - Bring back Vision Sessions test using InternVL2'。这表明需要修复会话中多模态输入的处理错误和测试回归。

实现拆解

实现分为三个部分：1) 在 scheduler_output_processor_mixin.py 的 process_batch_result_decode 函数中添加条件 req.session is None，防止清除会话中的多模态特征；2) 在 session_controller.py 的 create_req 函数中添加偏移调整逻辑，当 BOS 令牌被裁剪时，将 mm_item.offsets 每个元素减 1；3) 在测试文件 test_session_control.py 中，将模型从 'llava-onevision-qwen2-7b-ov' 改为 'OpenGVLab/InternVL2-2B'，并简化输入处理，直接使用文本而非 input_ids。

关键文件：

- python/sglang/srt/managers/scheduler_output_processor_mixin.py（模块 scheduling）：防止错误清除会话中的多模态特征，确保特征重用
- python/sglang/srt/managers/session_controller.py（模块 session management）：调整多模态偏移量以匹配 BOS 令牌裁剪，保证输入对齐
- test/registered/sessions/test_session_control.py（模块 testing）：恢复并更新视觉会话测试，验证修复效果

关键符号：process_batch_result_decode, create_req, test_session_control

评论区精华

Review 中没有详细讨论，只有 mickqian 的批准。但 Issue 评论中提到一个 follow-up PR #21501，用于在会话关闭时释放多模态特征并加强偏移调整，表明本PR是基础修复的一部分。

- 暂无高价值评论线程

风险与影响

- 风险：风险包括：偏移调整逻辑可能未覆盖所有情况，如 offsets 为空时；测试变更可能引入对 InternVL2 模型的依赖，影响其他测试；内存管理条件 req.session is None 可能不够全面，例如会话状态变化时。需要确保这些修改不会引入回归错误。
- 影响：对用户影响：修复后，多模态会话将正常工作，避免特征丢失或偏移错误。对系统影响：提高多模态处理的稳定性和测试覆盖。对团队影响：恢复视觉会话测试，有助于后续开发。
- 风险标记：偏移调整潜在错误，测试模型变更影响，内存管理条件不完整

关联脉络

- PR #21501 Release mm features on session close and support multiple /rerun-ut specs: 作为本 PR 的后续修复，扩展了多模态特征在会话关闭时的内存释放和偏移调整逻辑。