PR #21501 完整报告

sgl-project/sglang

Release mm features on session close and support multiple /rerun-ut specs

合并时间：2026-03-27 09:31

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21501

执行摘要

本 PR 修复了会话关闭时多模态特征未释放导致的 GPU 内存泄漏问题，通过统一清理模式并扩展 CI 的 /rerun-ut 命令支持多个测试规格。变更直接影响系统内存管理和 CI 效率，属于重要 bugfix，建议关注 session_controller.py 中的内存清理逻辑。

功能与动机

PR 旨在解决两个问题：首先，会话请求（session reqs）跳过正常清理路径，导致多模态特征张量在 GPU 内存中泄漏，这是 #21269 的后续跟进；其次，改进 CI 命令，使 /rerun-ut 能够一次性处理多个测试规格，如 /rerun-ut test_a.py test_b.py，以提升测试流程的灵活性。

实现拆解

实现分为两部分：

1. 多模态特征清理：
   - 在 MultimodalInputs 类（schedule_batch.py）中添加 release_features() 方法，遍历 mm_items 并将 feature 设为 None，统一释放 GPU 内存。
   - 在 session_controller.py 的 _close() 方法中，会话关闭时调用 release_features()，确保会话相关特征被释放；同时添加 BOS 偏移调整的防御性 clamp 和警告。
   - 在 scheduler.py 和 scheduler_output_processor_mixin.py 中，用 release_features() 替换原有的内联清理代码，减少重复逻辑。
2. CI 命令增强：
   - 重构 slash_command_handler.py，将 handle_rerun_ut 拆分为 _resolve_and_dispatch_ut，支持解析多个测试规格并分发工作流，最后整合结果到单个评论中。

评论区精华

无 review 评论，变更由作者直接合并，未经过外部技术讨论或争议。

风险与影响

风险：

- 内存释放可能不完整，例如在 session_controller.py 中通过 id(mm) 去重，若多模态输入结构复杂，可能导致遗漏释放。

- BOS 偏移调整的 clamp 逻辑（max(0, s - 1)）可能引入位置计算错误，影响多模态输入的语义正确性。
- CI 重构可能引入回归，如并发处理或权限检查漏洞，需依赖现有测试覆盖。 影响：
- 对系统：显著减少 GPU 内存泄漏风险，提升长时间会话的稳定性。
- 对团队：/rerun-ut 命令的增强简化测试流程，但需确保 CI 脚本的向后兼容性。

关联脉络

本 PR 是 #21269 的直接后续，延续了多模态输入会话修复的工作线；同时，与 #21495 相关，后者修复了 /rerun-ut 的并发问题，显示 CI 基础设施的持续演进。结合近期 PR 如 #21490（flush_cache 重构），可见团队在调度和内存管理方面的优化趋势。