PR #20782 完整报告

sgl-project/sglang

[MPS] Add StreamContext stub

合并时间：2026-03-26 11:59

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/20782

执行摘要

本 PR 通过向 _mps_stub.py 添加 StreamContext 存根，修复了 MacOS 上 MPS 后端启动时服务器崩溃的问题，是 MacOS 支持工作的一部分，变更简单但关键。

功能与动机

动机源于 Issue #20728，用户在 MacOS 上使用 MPS 设备启动 SGLang 服务器时遭遇崩溃。修复基于历史 PR 19549，旨在扩展系统对 MacOS 的兼容性。

实现拆解

变更仅涉及文件 python/sglang/_mps_stub.py。新增 StreamContext 类模拟 torch.cuda.StreamContext，方法均为空操作；在 install() 函数中将该类添加到 monkey-patch 列表，确保 MPS 后端初始化时注入。

评论区精华

review 中仅有 gemini-code-assist[bot] 的评论，指出实现合适且与现有存根一致，无争议；mickqian 直接批准。

风险与影响

风险较低，但存根需确保接口兼容性，且缺少单元测试可能掩盖问题。影响限于 MPS 用户，改善启动稳定性。

关联脉络

关联 Issue #20728 和 master issue #19137，是 MacOS 支持系列工作的一环；历史 PR 19549 为基础，但近期 PR 列表中无直接相关变更。