PR #25026 完整报告

sgl-project/sglang

[Bench] Add MEM profile activity to bench_serving

合并时间：2026-05-14 07:22

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25026

执行摘要

- 一句话：bench_serving 新增 MEM 内存分析选项
- 推荐动作：该 PR 逻辑简单，可直接合并。但建议未来考虑在 help 中明确 MEM 选项的依赖（如需要 CUDA 设备），并补充简单的单元测试确保 choices 包含 MEM。

功能与动机

PR body 中指出："Enables memory profiling alongside existing CPU/GPU/CUDA_PROFILER/XPU activities, so users can capture memory snapshots during serving benchmarks without separate tooling." 即让用户无需单独工具即可在基准测试中捕获内存快照。

实现拆解

1. 在 python/sglang/bench_serving.py 的 --profile-activities 参数的 choices 列表中添加 "MEM" 条目。
2. 同步更新 help 文本，说明 MEM 的作用："MEM dumps a torch.cuda.memory snapshot, viewable at https://pytorch.org/memory_viz"。
3. 无其他文件变更，后端的 MEM 处理逻辑由 torch.profiler 原生支持，无需额外代码。

关键文件：

- python/sglang/bench_serving.py（模块 基准测试；类别 source；类型 core-logic）：唯一变更文件，修改了 --profile-activities 参数的 choices 和 help 文本，新增 MEM 选项。

关键符号：未识别

关键源码片段

python/sglang/bench_serving.py

唯一变更文件，修改了 --profile-activities 参数的 choices 和 help 文本，新增 MEM 选项。

```
# --profile-activities 参数新增 MEM 选项
parser.add_argument(
    "--profile-activities",
    type=str,
    nargs="+",
```

```
    default=["CPU", "GPU"],
    # choices 新增 "MEM"，用于 torch.cuda.memory 快照
    choices=["CPU", "GPU", "CUDA_PROFILER", "XPU", "MEM"],
    help="Profiler activities to capture: CPU, GPU, XPU, CUDA_PROFILER, MEM "
    "(MEM dumps a torch.cuda.memory snapshot, viewable at https://pytorch.org/memory_viz).",
)
```

评论区精华

该 PR 仅有 1 条来自 gemini-code-assist[bot] 的 quota 警告，非技术性讨论。Qiaolin-Yu 直接审核通过，无 review 评论。因此没有实质性的技术讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。仅修改了 argparse 的 choices 枚举和 help 文本，不会影响现有逻辑。若用户误用 MEM（例如在不支持 CUDA 的环境下），torch.profiler 可能抛出错误，但该行为与现有错误处理一致，不构成新增风险。
- 影响：影响范围有限。仅影响使用 bench_serving 且指定 --profile-activities MEM 的用户，其他用户完全无感知。为调试 GPU 内存泄漏或峰值内存使用的开发者提供了便利。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR