

PR #34423 完整报告

sgl-project/sglang

[diffusion] fix: nightly diffusion benchmark passes the retired --warmup flag

合并时间: 2026-08-12 10:25

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/34423>

执行摘要

- 一句话: 修复 diffusion 基准误传弃用 --warmup 标志
- 推荐动作: 值得快速浏览, 尤其是 CI 配置维护者。关键设计决策是用 --warmup-mode server 替代而不是删除标志, 避免静默关闭预热; 作者提出的 pre-merge 参数预检 lint 是低成本的后续改进, 建议纳入 CI 治理。

功能与动机

PR 正文指出 nightly-test-diffusion 每次运行都失败, 所有 sglang server 在参数解析阶段报错 `sglang: error: ambiguous option: --warmup could match --warmup-mode, --warmup-resolutions, --warmup-steps`, 随后服务器立即退出, harness 却要等满 2400s 健康检查超时才继续, 约 40 分钟 / 用例只产出失败。只删除标志也不行, 因为 warmup_mode 默认 None, _adjust_warmup 会解析为 off, 会让 eager 用例悄悄失去预热, 产生静默基准回归。

实现拆解

1. 修改 scripts/ci/utls/diffusion/comparison_configs.json: 将 Wan、LTX-2、Cosmos 等 11 个 benchmark 用例的 serve_args 全部由 --warmup 改为 --warmup-mode server, 保持原有启动时预热内核再进入健康检查的语义。
2. 修改 scripts/ci/utls/diffusion/run_comparison.py: 更新 wait_for_health 调用前的注释, 说明现在 sglang 通过 --warmup-mode server 在 health check 前完成合成请求预热。
3. 验证: 作者在 H200 上用 ServerArgs.add_cli_args 逐条解析全部 11 个 serve_args, 全部通过; 再注入旧 --warmup 做阴性对照, 能稳定复现 ambiguous option 报错, 证明检查有效。
4. 后续改进 (本次未做): 作者建议在 pre-merge lint 中解析该配置的所有 serve_args, 成本极低, 但能当天捕获此类 flag 弃用; 明确留在本 PR 之外保持改动最小。

关键文件:

- scripts/ci/utls/diffusion/comparison_configs.json (模块 CI 脚本; 类别 infra; 类型 configuration): 核心修复文件: 11 条 serve_args 全部去掉已弃用的 --warmup, 改为 --warmup-mode server, 否则 nightly diffusion 基准每次都在参数解析处崩溃。
- scripts/ci/utls/diffusion/run_comparison.py (模块 CI 脚本; 类别 infra; 类型 documentation): 同步更新 wait_for_health 调用处的注释, 避免后续维护者再次传入旧

标志。

关键符号：未识别

评论区精华

本 PR 没有 review 评论。最有价值的讨论来自 PR 正文的作者自述：他提出把解析所有 `serve_args` 做成 pre-merge lint，认为用 `ServerArgs.add_cli_args` 校验成本极低，却能在 flag 弃用当天就发现问题。这个方向值得 CI 治理跟进，但目前仍未落地。

- 暂无高价值评论线程

风险与影响

- 风险：风险主要集中在 CI 配置层：`comparison_configs.json` 若未来新增条目仍可能再次漏掉 `--warmup-mode`；目前只在 H200 单机做过参数解析验证，未覆盖多平台；若 `_adjust_warmup` 的默认解析逻辑变化，基准仍可能静默退化。由于不涉及运行时代码，对用户请求路径无风险。
- 影响：修复 `nightly-test-diffusion` 的持续失败，消除约 40 分钟 / 用例的无效等待，恢复该基准任务的产出。影响范围仅限 `scripts/ci` 下的 `diffusion` 基准配置与注释，不触碰任何模型运行时或用户请求路径，对团队 CI 维护者是明确的收益。
- 风险标记：依赖已弃用参数，缺少自动化防护

关联脉络

- 暂无明显关联 PR