# PR #26120 完整报告

sgl-project/sglang

[diffusion] CI: disable torch compile in nightly comparison

合并时间：2026-05-22 23:36

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26120

## 执行摘要

- 一句话：禁用 diffusion 夜间对比中的 torch compile
- 推荐动作：可快速合入的维护性变更。建议关注后续是否有 torch compile 兼容性修复的 PR，届时可恢复此选项。

## 功能与动机

夜间对比 CI 作业中 Wan A14B warmup 超时（1200s 限制），日志显示 torch/_inductor/select_algorithm.py 中 Triton 资源错误。移除 --enable-torch-compile 以规避问题并恢复对比稳定性。

## 实现拆解

1. 修改配置文件：在 scripts/ci/utils/diffusion/comparison_configs.json 中，对所有 SGLang 模型的 serve_args 字符串统一删除 --enable-torch-compile 选项（共 10 处），保留 --warmup 及其他参数。
2. 更新超时注释：在 scripts/ci/utils/diffusion/run_comparison.py 中将 HEALTH_TIMEOUT 的注释从“40 min — FLUX.2-dev needs ~10 min download + torch.compile”改为“40 min — keep large model download/warmup headroom”，避免关于 torch compile 的误导。

关键文件：

- scripts/ci/utils/diffusion/comparison_configs.json（模块 CI 配置；类别 infra；类型 infrastructure）：核心配置文件，移除了所有 diffusion 模型的 torch compile 启动参数，直接修复超时问题。
- scripts/ci/utils/diffusion/run_comparison.py（模块 CI 脚本；类别 infra；类型 infrastructure）：更新了健康检查超时注释，移除误导性的 torch compile 说明。

关键符号：未识别

## 评论区精华

无 review 讨论。

- 暂无高价值评论线程

## 风险与影响

- 风险：低风险。仅禁用了加速编译功能，不影响模型正确性；移除 torch.compile 后 warmup 时间缩短，但延迟性能可能略有下降。CI 对比仍能正常运行。
- 影响：直接影响 diffusion 夜间对比 CI 的 SGLang 用例，避免因 torch compile 导致的超时失败。用户无感知。
- 风险标记：暂无

## 关联脉络

- PR #26044 ci: guard diffusion gt publishing: 同为 diffusion CI 稳定性改进，均涉及 diffusion 发布 / 对比流程。