PR #22251 完整报告

sgl-project/sglang

[diffusion] CI: fix consistency check

合并时间：2026-04-07 23:43

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22251

执行摘要

本 PR 修复了扩散模型 CI 中的一致性检查问题，通过移除不稳定的 'sana_image_t2i' 测试用例、调整阈值配置和优化 CI 工作流，旨在提升测试稳定性和可靠性，减少 flaky 测试对开发流程的干扰。

功能与动机

动机源自测试中的波动问题，特别是 'sana' 和 'mova' 案例。作者在 Issue 评论中表示 ' TODO:

1. investigate the reason of fluctuation in sana, mova cases'，本 PR 通过调整测试套件来缓解问题，确保 CI 的一致性检查更健壮，避免因阈值过严或用例不稳定导致的 CI 失败。

实现拆解

- CI 工作流（.github/workflows/diffusion-ci-gt-gen.yml）：添加环境变量 SGLANG_IS_IN_CI 和 SGLANG_CUDA_COREDUMP，优化生成输出和发布 GT 图像的步骤，新增 B200 GPU 测试支持，确保测试环境一致。
- 测试脚本（gen_diffusion_ci_outputs.py）：引入 _maybe_pin_update_weights_model_pair 函数，用于更新模型权重，提升测试可重复性。
- 阈值配置（consistency_threshold.json）：删除 'sana_image_t2i' 的阈值条目，因为该测试用例已在配置中移除，避免无效配置。
- 测试用例配置（testcase_configs.py）：移除 'sana_image_t2i' 用例，并为 flux_2_nvfp4_t2i 和 ltx_2_two_stage_t2v 启用一致性检查（通过删除 run_consistency_check=False），简化测试逻辑。
- 测试逻辑（test_server_common.py）：修改 run_and_collect 函数，添加 collect_perf 参数以在生成 GT 时跳过性能收集，移除 LoRA 相关的后端逻辑，代码如下：python def run_and_collect(ctx: ServerContext, case_id: str, generate_fn: Callable[[str, openai.Client], tuple[str, bytes]], collect_perf: bool = True,) -> tuple[RequestPerfRecord | None, bytes]:

评论区精华

Review 中仅有的评论来自 gemini-code-assist[bot]，指出：

> 'Enabling the consistency check for flux_2_nvfp4_t2i without adding a corresponding entry in consistency_threshold.json may cause CI failures due to

```
strict default thresholds.'
```

这提示了阈值管理的设计权衡：启用检查需要相应配置，而本 PR 通过移除不稳定用例间接解决了问题，但未直接回应 bot 的建议，可能隐含了对测试波动性的妥协。

风险与影响

- 风险：移除 'sana_image_t2i' 测试用例可能掩盖潜在模型问题；删除阈值配置可能导致其他测试误判；CI 环境变量变更可能引入新的不稳定性；修改 collect_perf 逻辑可能影响性能监控。
- 影响：对最终用户无直接功能影响，但提升了 CI 的稳定性和团队开发效率；测试覆盖略有减少，聚焦于更可靠的案例；系统级影响限于测试基础设施，程度中等。

关联脉络

本 PR 与历史 PR #15236（'[CI] Add consistency test in CI'）紧密相关，后者引入了扩散模型的一致性测试。当前修复处理了其中发现的波动问题，如 sana 案例的不稳定性，反映了测试套件在迭代中的优化过程，展现了持续集成中测试稳定性的演进趋势。