PR #22854 完整报告

sgl-project/sglang

[diffusion] CI: reset thresholds

合并时间：2026-04-15 21:11

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22854

执行摘要

- 一句话：重置扩散模型 CI 性能基准阈值，更新基准数据以匹配 H100 运行结果。
- 推荐动作：建议：对于维护扩散模型 CI 的工程师，此 PR 值得关注基准数据的更新逻辑和容差调整策略；对于其他开发者，可了解如何通过 CI 脚本优化错误处理。

功能与动机

根据提交历史，PR 旨在 'reset thresholds' 和 'refresh baselines'，以使用 2026-04-15 的 H100 CI 运行数据更新性能基准，确保 CI 测试的准确性和最新性。

实现拆解

1. 更新性能基准文件：修改 python/sglang/multimodal_gen/test/server/perf_baselines.json，更新元数据（如描述和最后更新时间）、容差值（如将 long_term.e2e 从 0.1 改为 0.15）和场景性能数据（如 DenoisingStage 从 14289.46 ms 降至 12404.23 ms），以匹配新基准。
2. 增强 CI 脚本错误处理：修改 scripts/ci/utils/diffusion/run_comparison.py，使 run_comparison 函数返回输出数据，并在有错误时打印错误信息并退出非零（sys.exit(1)），提高 CI 的健壮性。
3. 修正配置任务类型：修改 scripts/ci/utils/diffusion/comparison_configs.json，将任务 id 为 ltx2.3_twostage_ti2v_2gpus 的 task 从 "image-to-video" 改为 "text-image-to-video"，并添加 reference_image、width、height 参数，确保配置准确性。

关键文件：

- python/sglang/multimodal_gen/test/server/perf_baselines.json（模块 性能基准；类别 test；类型 test-coverage）：性能基准核心文件，更新了所有扩散场景的容差和阶段时间数据，直接影响 CI 测试通过标准。
- scripts/ci/utils/diffusion/run_comparison.py（模块 CI 脚本；类别 infra；类型 infrastructure）：CI 比较脚本，增强错误处理逻辑，在运行时错误时退出非零，提升 CI 健壮性。
- scripts/ci/utils/diffusion/comparison_configs.json（模块 CI 配置；类别 infra；类型 infrastructure）：CI 比较配置文件，修正任务类型和参数，确保测试场景准确性。

关键符号：未识别

关键源码片段

python/sglang/multimodal_gen/test/server/perf_baselines.json

性能基准核心文件，更新了所有扩散场景的容差和阶段时间数据，直接影响 CI 测试通过标准。

```
{
  "metadata": {
    "model": "Diffusion Server",
    "hardware": "CI H100 80GB pool",
    "description": "Reference numbers captured from CI H100 runs (2026-04-15).", //
    更新描述，添加具体日期
    "last_updated": "2026-04-15" // 新增字段，记录最后更新时间
  },
  "tolerances": {
    "long_term": {
      "e2e": 0.15, // 从容差从0.1提高到0.15，放宽端到端测试标准
      "denoise_stage": 0.1, // 去噪阶段容差从0.05提高到0.1
      "non_denoise_stage": 0.5, // 非去噪阶段容差从0.4提高到0.5
      "denoise_step": 0.25, // 去噪步从容差从0.2提高到0.25
      "denoise_agg": 0.15 // 去噪聚合容差从0.1提高到0.15
    },
    "pr_test": {
      "e2e": 0.25, // PR测试容差从0.2提高到0.25
      "denoise_stage": 0.25, // 从0.2提高到0.25
      "non_denoise_stage": 0.8, // 保持不变
      "denoise_step": 0.3, // 从0.25提高到0.3
      "denoise_agg": 0.2 // 从0.15提高到0.2
    }
  },
  // ... 其他部分如场景数据省略
}
```

scripts/ci/utils/diffusion/run_comparison.py

CI 比较脚本，增强错误处理逻辑，在运行时错误时退出非零，提升 CI 健壮性。

```
def main():
    # ... 前略代码（加载配置等）
    print(f"Loaded {len(config['cases'])} comparison case(s) from {args.config}")

    output_data = run_comparison( # 修改：将函数调用赋值给变量以获取返回数据
        config=config,
        case_ids=args.case_ids,
        frameworks=args.frameworks,
        # ... 其他参数
        dry_run=args.dry_run,
    )

    # 新增：检查输出数据中的错误并退出非零
    errors = [r for r in output_data.get("results", []) if r.get("error")]
```

```
    if errors and not args.dry_run:
        print(f"\n{len(errors)} case(s) had errors:")
        for e in errors:
            print(f"  {e['case_id']} ({e['framework']}): {e['error']}")
        sys.exit(1) # 有错误时退出非零状态码

if __name__ == "__main__":
    main()
```

scripts/ci/utils/diffusion/comparison_configs.json

CI 比较配置文件，修正任务类型和参数，确保测试场景准确性。

```
{
    "id": "ltx2.3_twostage_ti2v_2gpus",
    "model": "Lightricks/LTX-2.3",
    "task": "text-image-to-video", // 从 "image-to-video" 改为 "text-image-to-video"，修正任务类型
    "prompt": "The cat starts walking slowly towards the camera.",
    "reference_image": true, // 新增字段，启用参考图像
    "width": 768, // 新增宽度参数
    "height": 512, // 新增高度参数
    "num_frames": 121,
    "seed": 42,
    "num_gpus": 2
}
```

评论区精华

Review 评论为空，无讨论。作者在关联 Issue 中触发了 CI 测试（/tag-and-rerun-ci），表明变更已通过 CI 验证。

- 暂无高价值评论线程

风险与影响

- 风险：风险包括：基准数据更新可能导致现有测试通过性变化，如果新基准不准确可能掩盖性能回归；CI 脚本错误处理变更（如退出非零）可能引入新的失败条件，影响 CI 流程。具体在 perf_baselines.json 中，容差值调整（如 pr_test.e2e 从 0.2 提高到 0.25）可能放松测试标准，需确保阈值合理。
- 影响：影响范围：扩散模型 CI 测试的通过率和准确性；对用户无直接影响，但确保团队在性能回归检测上的可靠性。影响程度：中等，因为基准更新会影响所有扩散模型相关 CI 测试，但仅限于测试和基础设施层面。
- 风险标记：基准数据变更，CI 脚本行为改变

关联脉络

- PR #22810 [diffusion] CI: refactor diffusion ci and reduce redundancy: 同为扩散模型 CI 相关重构，可能共享代码或测试逻辑，提供上下文参考。