

PR #29767 完整报告

sgl-project/sglang

[CI] Relax ModelOpt NVFP4 diffusion consistency thresholds

合并时间: 2026-07-01 14:45

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29767>

执行摘要

本 PR 为 ModelOpt NVFP4 量化的扩散模型 (flux2 和 wan22) 在 B200 CI 中的一致性测试添加了专有阈值, 解决了因默认阈值过紧导致的 CI 失败。仅修改测试配置文件, 无生产代码变更, 风险极低。

功能与动机

PR #29315 的 base CI 运行中, `flux2_modelopt_nvfp4_t2i` 和 `wan22_modelopt_nvfp4_t2v` 两个 case 的一致性指标未通过默认阈值。例如 flux2 的 ssim 实测 0.8864 低于默认的 0.95, psnr 实测 17.4954 低于默认的 28.0。作者 BBuf 在 PR body 中列出了详细的实测值与旧阈值对比, 表明需要根据实际量化精度放宽阈值。

实现拆解

1. 识别失败 case: 从 CI 日志中锁定两个失败 case 及其具体指标。
2. 添加专有阈值: 在 `python/sglang/multimodal_gen/test/server/consistency_threshold.json` 的 `cases` 字典中新增两个条目, 每个条目包含 `clip_threshold`、`ssim_threshold`、`psnr_threshold`、`mean_abs_diff_threshold` 四个阈值。阈值设定略低于实测值, 确保通过。
3. 保持全局缺省不变: 其他 case 和全局阈值 (在测试框架中定义) 不受影响, 保证风险隔离。
4. 验证: 通过本地 `git diff --check` 和逻辑比对确认新阈值能允许实测值通过。

`python/sglang/multimodal_gen/test/server/consistency_threshold.json`

唯一变更文件, 为两个扩散模型 case 添加了专有的一致性测试阈值。

```
// python/sglang/multimodal_gen/test/server/consistency_threshold.json
// 新增 flux2_modelopt_nvfp4_t2i 和 wan22_modelopt_nvfp4_t2v 两个 case 条目
{
  "_comment": "Some cases use lower thresholds; raise them if quality/perf improves later.",
  "cases": {
    // ... 其他 case ...
    // 新增: flux2 图像生成, NVFP4 量化, 阈值基于实际 CI 失败观测值
    "flux2_modelopt_nvfp4_t2i": {
      "clip_threshold": 0.92,
      "ssim_threshold": 0.86, // 原默认 0.95, 实测 0.8864
      "psnr_threshold": 17.0, // 原默认 28.0, 实测 17.4954
      "mean_abs_diff_threshold": 9.5 // 原默认 8.0, 实测 8.7758
    },
  },
}
```

```
// ... 其他 case ...
// 新增: wan22 视频生成, NVFP4 量化
"wan22_modelopt_nvfp4_t2v": {
  "clip_threshold": 0.90, // 原视频默认 0.90, 实测 0.9978
  "ssim_threshold": 0.90, // 原视频默认 0.92, 实测 0.9097
  "psnr_threshold": 24.0, // 原视频默认 24.0, 实测 32.0379
  "mean_abs_diff_threshold": 10.0 // 原视频默认 10.0, 实测 4.3456
},
"fastwan2_2_ti2v_5b": {
  "clip_threshold": 0.90,
  "ssim_threshold": 0.88,
  "psnr_threshold": 24.0,
  "mean_abs_diff_threshold": 10.0
}
// ... 其余 case ...
}
```

评论区精华

无明显 review 讨论。PR body 中作者 BBuf 提供了详细的失败指标和阈值决策依据。

风险与影响

- 风险：低风险。仅修改测试配置，无生产代码变更。新阈值基于实测值，但可能略低于理想精度；未来量化质量提升后可考虑收紧。
- 影响：解决 B200 CI 中特定扩散模型的持续失败，保持 CI 主干健康度。不影响其他硬件或测试 case。

关联脉络

- 触发 PR #29315：该 PR 的 CI 运行首先暴露了阈值过紧问题。
- 无其他历史 PR 关联：仅独立的测试维护。