

PR #29824 完整报告

sgl-project/sglang

[diffusion] CI: tighten multimodal-gen consistency thresholds

合并时间: 2026-07-02 01:32

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29824>

执行摘要

本 PR 基于最新 main 分支成功的多模态生成 CI 运行，系统性收紧 H100 上所有已有 test case 的一致性阈值 (clip、ssim、psnr、mean_abs_diff)，采用固定 margin 公式。这是一次纯粹的测试配置优化，无源码逻辑变更，旨在提高 CI 对模型退化检测的灵敏度，降低假阴性概率。

功能与动机

根据 PR body 所述，最近一次 main 分支 scheduled multimodal-gen run (run 27242606028) 完成后，作者希望利用该运行的实际观测指标，收紧现有threshold的容忍度。规则设定为：

- clip_threshold: latest worst clip 减 0.02，上限 0.98
- ssim_threshold: latest worst SSIM 减 0.04，上限 0.95
- psnr_threshold: latest worst PSNR 减 3 dB，上限 30 dB
- mean_abs_diff_threshold: latest worst mean abs diff 加 3，下限 4

所有调整均保留合理 margin，避免过度敏感。不改变全局默认值，不新增 case。

实现拆解

本 PR 只改动一个文件 `python/sglang/multimodal_gen/test/server/consistency_thresholds/h100.json`，内容是完全的 JSON 配置。变更过程：

1. 从 CI 产物中提取所有 case 在 run 27242606028 中的最差观测值。
2. 对每个 case 的四个指标分别应用上述公式计算新阈值。
3. 确保结果不超出 cap/floor 限制，且仅收紧现有 override，未引入新条目。
4. 使用 `python3 -m json.tool` 验证 JSON 格式，`git diff --check` 检查空白。

由于是纯配置调整，不需要运行单元测试（但本地因环境缺失 pybase64 跳过）。

`python/sglang/multimodal_gen/test/server/consistency_thresholds/h100.json`

唯一变更文件，调整了所有 20+ 个测试用例的 4 个一致性指标阈值，基于最新 CI 运行数据。这是提升测试灵敏度的核心操作。

```
{  
    // 示例: ideogram4_nvfp4_t2i 的收紧前后对比
```

```
// ( 注释遵守盘古排版: 中文与英文、数字之间留空格 )
"clip_threshold": 0.98, // 从 0.97 收紧 ( 减 0.02 后 cap 0.98)
"ssim_threshold": 0.95, // 从 0.78 大幅收紧 ( 减 0.04 后 cap 0.95)
"psnr_threshold": 18.0, // 保持不变 ( 历史观测值未触发减小 )
"mean_abs_diff_threshold": 4.0 // 从 18.0 断崖式收紧 ( 加 3 后 floor 4)
}
```

评论区精华

本 PR 无人类 reviewer 评论, 仅 Gemini Code Assist 机器人自动回复, 无额外建议。

风险与影响

- 风险: 新阈值基于单次 CI 运行, 若有偶发性波动, 可能导致不必要的 CI 失败。但 margin (减 0.02/0.04/3 dB 或加 3) 预留了一定缓冲。若模型质量因合法优化提升, 阈值不会自动放宽, 需手动更新 (但这与本次收紧方向一致, 不算问题)。
- 影响: 直接影响 diffusion 测试套件, 使得 CI 对图像 / 视频生成质量的微小退化更敏感。开发者在提交与多模态生成相关的代码时, 需关注新阈值, 失败时确认是否真实退化。无性能或业务逻辑影响。

关联脉络

本 PR 与 #29791 (添加 5090 diffusion GPU CI 防护) 共同构成 diffusion CI 增强系列, 一个扩展硬件覆盖, 一个提升测试可靠性。此前还有 #29789 清理 diffusion 死代码, #29693 为 AMD 添加 diffusion JIT 测试。整体反映团队正在系统性地加固多模态生成测试基础设施, 为后续模型迭代提供更稳定的质量门。