PR #28778 完整报告

sgl-project/sglang

[diffusion] CI: remove flaky layerwise offload diffusion case

合并时间：2026-06-20 17:27

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28778

执行摘要

- 一句话：移除不稳定的 layerwise_offload 扩散测试用例
- 推荐动作：该 PR 是清理 flaky 测试的常规维护，不值得精读。但如果团队维护 layerwise offload 功能，建议关注是否用更小、更快的模型重建测试以保持回归检测，或依赖其他扩散测试场景来覆盖。

功能与动机

PR body：该 case 是某些 runner 上的 flaky perf gate；ZImage coverage 已通过 zimage_image_t2i 和相关用例覆盖；删除独立的 --dit-layerwise-offload 场景而不是放宽阈值。

实现拆解

1. 删除测试用例注册：在 gpu_cases.py 的测试列表中去掉 DiffusionTestCase("layerwise_offload", ...) 及相关的 TODO 注释。
2. 移除性能基线：从 perf_baselines.json 中删除整个 layerwise_offload 键，它包含 stages_ms、denoise_step_ms、expected_e2e_ms 等约 70 行数据。
3. 移除一致性阈值：从 consistency_threshold.json 中删除 layerwise_offload 条目（clip_threshold 等四个阈值）。
4. 移除组件精度跳过：从 accuracy_config.py 的 SKIP_COMPONENTS 中删除 layerwise_offload 映射（VAE、TRANSFORMER、TEXT_ENCODER 三个组件的跳过原因）。
5. 伴随基线调整：在 perf_baselines.json 中微调了 fsdp-inference 的 step 3 denoise 时间（257.61 ms → 650.0 ms）和 expected_e2e_ms（2775.88 → 2850.0），以及另一条目的 expected_avg_denoise_ms（500.63 → 520.0），这些来自最后一笔 commit 的 flaky 2-gpu 基线放松。

关键文件：

- python/sglang/multimodal_gen/test/server/perf_baselines.json（模块 性能基线；类别 test；类型 test-coverage）：删除了 layerwise_offload 的完整性能基线块（约 70 行），并微调了 fsdp-inference 等条目的基线值。
- python/sglang/multimodal_gen/test/server/gpu_cases.py（模块 测试用例；类别 test；类型 test-coverage）：移除了 layerwise_offload 测试用例注册以及相关的 TODO 注释，是该 PR 的核心删除点。

- python/sglang/multimodal_gen/test/server/accuracy_config.py（模块 精度配置；类别 test；类型 test-coverage）：从 SKIP_COMPONENTS 字典中移除了 layerwise_offload 条目，该条目定义了三个组件的跳过原因。
- python/sglang/multimodal_gen/test/server/consistency_threshold.json（模块 一致性阈值；类别 test；类型 test-coverage）：删除了 layerwise_offload 的一致性阈值条目（clip_threshold 等四个指标）。

关键符号：未识别

关键源码片段

python/sglang/multimodal_gen/test/server/gpu_cases.py

移除了 layerwise_offload 测试用例注册以及相关的 TODO 注释，是该 PR 的核心删除点。

```
# gpu_cases.py – 删除了 `layerwise_offload` 测试用例及其 TODO 注释
# head 版本中，在 `flux_2_klein_base_image_t2i` 之后直接是 `zimage_image_t2i`
# 此前（base）定义如下（已删除）：
# DiffusionTestCase(
# "layerwise_offload",
# DiffusionServerArgs(
# model_path=DEFAULT_SMALL_MODEL_NAME_FOR_TEST,
# dit_layerwise_offload=True,
# dit_offload_prefetch_size=2,
# ),
# ),
# 同时删除的 TODO 注释：
# # TODO: replace with a faster model to test the --dit-layerwise-offload
# # TODO: currently, we don't support sending more than one request...
```

评论区精华

该 PR 没有 review 评论，仅有一条 gemini-code-assist 的自动配额警告和作者的重跑指令，无实质性技术讨论。

- 暂无高价值评论线程

风险与影响

- 风险：主要风险是移除了对 --dit-layerwise-offload 功能的 CI 回归检测。虽然该功能有间接覆盖（如 zimage_image_t2i 也使用了类似的 offload 机制），但专门的 layerwise offload 场景不再被 CI 验证，未来重构该功能时可能遗漏退化。不过该测试本身就是 flaky，移除后 CI 稳定性提高，权衡后风险可接受。
- 影响：对用户无直接功能影响。对 CI 系统：减少一次 flaky 失败源，略微缩短跑 CI 总时间。对开发团队：需要意识到 layerwise offload 的 CI 覆盖已移除，未来若需加强验证可考虑用更小的模型或更稳健的基线重建测试。
- 风险标记：移除专用测试覆盖，伴随基线微调

关联脉络

- PR #28773 [Diffusion] Keep FastHunyuan VAE resident on high-memory GPUs: 同样属于 diffusion 模块的 offload 策略优化，该 PR 移除了专门的 layerwise offload 测试，而 28773 增强了 VAE 驻留行为，两者在 diffusion 内存管理逻辑上有间接关联。