PR #29434 完整报告

sgl-project/sglang

[diffusion] nightly: track SGLang-Diffusion only

合并时间：2026-06-28 14:23

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/29434

执行摘要

- 一句话：夜间测试改为仅跟踪 SGLang-Diffusion 并更新用例
- 推荐动作：值得相关部门（Diffusion 开发、CI 团队）精读。本文展示了从跨框架比较转为自跟踪的决策过程，以及 warmup 策略的实践经验（revert 和重新实现）。尤其是 warmup 与延迟指标的选择，对于确保 CI 结果可靠性和一致性有参考价值。

功能与动机

The diffusion nightly test now tracks SGLang-Diffusion itself (regression over time) rather than comparing against other serving frameworks. This PR drops the cross-framework framing and refreshes the benchmark case list.

实现拆解

1. 重命名 CI 任务

在 .github/workflows/nightly-test-nvidia.yml 中将 nightly-test-diffusion-comparison 改为 nightly-test-diffusion，同步修改 dispatch 选项、job_filter 条件以及 summary job 的 needs 依赖列表。

2. 更新 Benchmark 配置

在 scripts/ci/utils/diffusion/comparison_configs.json 中移除 ltx2_twostage_t2v（LTX-2），保留并升级 ltx2.3_twostage_ti2v_2gpus（调整为 --cfg-parallel-size 2），新增 ideogram4_fp8_t2i_2gpu（Ideogram-4，--tp-size 2 --attention-backend fa）和 cosmos3_super_t2v_2gpu（Cosmos3-Super，启用 SGLANG_DISABLE_COSMOS3_GUARDRAILS）。

3. 调整脚本文案与延迟报告

在 run_comparison.py 中修改模块 docstring，移除“cross-framework”字样；将 send_image_request_sglang、send_video_request_sglang 等函数的返回值从 server_latency 改为 client_latency，并在日志中标记 server latency 为诊断信息。在 generate_diffusion_dashboard.py 和 publish_comparison_results.py 中更新标题和 argparse 描述。

4. 优化 Warmup 策略（提交历史演进）

最初尝试完全依赖服务器内置 warmup 并删除客户端 warmup 循环，但发现服务器默认 warmup_steps=1 导致延迟回归。最终采用服务器 warmup 加上 --warmup-resolutions 参数匹配测试分辨率，同时保留 *e2e* 客户端延迟作为主指标。

关键文件：

- scripts/ci/utils/diffusion/run_comparison.py（模块 CI 脚本；类别 infra；类型 infrastructure；符号 send_image_request_sglang, send_video_request_sglang, send_image_conditioned_request_sglang, _build_sglang_cmd）：核心 benchmark 脚本，修改了 docstring、延迟报告方式（server→client），并增加了 warmup 相关注释。
- scripts/ci/utils/diffusion/comparison_configs.json（模块 配置数据；类别 infra；类型 configuration）：Benchmark 用例配置文件，移除 LTX-2，新增 Ideogram-4 和 Cosmos3-Super，调整 LTX-2.3 参数。
- .github/workflows/nightly-test-nvidia.yml（模块 工作流配置；类别 infra；类型 configuration）：CI 工作流定义，重命名 diffusion job 并同步所有引用点。
- scripts/ci/utils/diffusion/generate_diffusion_dashboard.py（模块 CI 脚本；类别 infra；类型 infrastructure；符号 generate_dashboard, main）：Dashboard 生成脚本，更新标题和描述。
- scripts/ci/utils/diffusion/publish_comparison_results.py（模块 CI 脚本；类别 infra；类型 infrastructure；符号 main）：结果发布脚本，更新描述文字。

关键符号：send_image_request_sglang, send_video_request_sglang, send_image_conditioned_request_sglang, _build_sglang_cmd

关键源码片段

scripts/ci/utils/diffusion/run_comparison.py

核心 benchmark 脚本，修改了 docstring、延迟报告方式（server→client），并增加了 warmup 相关注释。

```
# send_image_request_sglang 中的延迟报告修改
if perf_dump_path:
    server_latency = _read_perf_dump(perf_dump_path)
    if server_latency is not None:
        # 改为客户端 e2e 延迟为主指标，服务器侧延迟仅作为诊断信息
        print(
            f" Image generated in {client_latency:.2f}s (client e2e; "
            f"server-side {server_latency:.2f}s, diagnostic)"
        )
    return client_latency # 原返回 server_latency

# 类似修改在 send_video_request_sglang 和 send_image_conditioned_request_sglang 中
```

评论区精华

该 PR 无显式 review 评论。但从 7 个提交历史可见 warmup 策略的显著演进：

- 提交 fe3ac0b 尝试完全依赖服务器内置 warmup，删除客户端 warmup 循环。

- 提交 afa1c21 回退该修改，因为实测延迟翻倍（服务器 warmup_steps=1 不足）。
- 提交 8e54cfd 结合 --warmup 修复后重新采用服务器 warmup，并新增 --warmup-resolutions 精确匹配测试形状。
- 最终确认使用客户端 *e2e* 延迟作为报告指标，server latency 仅作诊断。
- 暂无高价值评论线程

风险与影响

- 风险：主要风险：
  - CI 任务重命名同步遗漏：nightly-test-nvidia.yml 中 workflow_dispatch 的 default、job_filter 的 if 条件以及 summary job 的 needs 列表必须全部一致，否则 workflow 可能无法正确触发或状态汇总失败。当前修改已全部覆盖。
  - 新模型配置错误：comparison_configs.json 中 ideogram4_fp8_t2i_2gpu 和 cosmos3_super_t2v_2gpu 的 serve_args 和 extra_env 需与实际模型兼容。Ideogram-4 使用 FP8 权重、TP2 和 FA backend；Cosmos3-Super 禁用 guardrails。配置经验证通过了 11 个用例的解析。
  - Warmup 变化导致延迟偏差：服务器 warmup 步骤数（默认 1 步）可能不足以充分预热 CUDA graph，但通过 --warmup-resolutions 精确匹配分辨率提升了效果。Client *e2e* 延迟作为主要指标更贴近用户感知，但与历史数据对比时需注意基线变化。
  - 无测试覆盖：此修改仅涉及 CI 基础设施，没有配套单元测试，但已有 PR CI 运行验证语法正确性。
  - 影响：影响范围：仅限 Diffusion nightly CI 测试流程。对用户：无终端用户影响。对系统：夜间测试用例从 9 个变为 11 个（移除 1 个，新增 2 个），测试执行时间可能略有增加。延迟指标从 server-side 改为 client-side *e2e*，历史 dashboards 数据不再完全可比，但 dashboard 标题和趋势图已更新以反映新基线。对团队：Diffusion 团队现在能更清晰地追踪 SGLang-Diffusion 自身的性能回归，不再受跨框架比较带来的噪声干扰。
- 风险标记：CI 任务重命名需同步多处，warmup 策略变更可能影响延迟数据，新模型配置可能未充分验证

关联脉络

- PR #29514 [diffusion] fix --warmup silently downgrading server-based warmup to request mode: 修复了 --warmup 静默降级问题，为本 PR 采用服务器 warmup 策略提供了基础。
- PR #28624 [diffusion] optimize LTX2.3 CFG/SP paths: 优化了 LTX2.3 性能，本 PR 更新了 LTX2.3 的 benchmark 配置（--cfg-parallel-size 2）。
- PR #29464 Fix EAGLE draft hidden dim extraction and centralize spec helpers: 同仓库最近合并的 PR，与本 PR 无直接关联，但展示了社区活跃度。