PR #22086 完整报告

sgl-project/sglang

[diffusion] CI: improve diffusion comparison benchmark setting for realistic perf and auto-discover ut

合并时间：2026-04-04 23:20

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22086

执行摘要

本 PR 主要针对 SGLang 仓库的扩散模型 CI 基准测试进行改进：通过增加推理步数获取更真实的性能数据，强制使用严格端口防止 CI 端口冲突，并重构端口逻辑以提升服务器启动稳定性；同时，自动发现单元测试文件以简化测试维护。这些变更增强了 CI 的可靠性和性能基准的准确性。

功能与动机

改进动机源于 PR body 中所述：Wan2.2 模型的基准测试中，num_inference_steps 仅为 2 导致去噪阶段失真，且双 DiT 模型的第二模型未能充分测试。增加步数至 20 可更真实地反映性能。此外，强制 --strict-ports 旨在避免端口误路由问题，确保 CI 服务器启动稳定。

实现拆解

- 基准测试配置优化：在 scripts/ci/utils/diffusion/comparison_configs.json 中，将 Wan2.2 用例的推理步数从 2 增加至 20，移除了不必要的参数覆盖，并添加了 LTX-2 TwoStage 基准测试用例，使用模型默认参数。
- 端口逻辑重构：在 python/sglang/multimodal_gen/runtime/server_args.py 中，提取 _require_port 方法统一端口验证，解决 master_port 为 None 时的崩溃问题，并设置默认值 30005。
- 单元测试自动发现：修改 python/sglang/multimodal_gen/test/run_suite.py，引入 _discover_unit_tests 函数自动发现单元测试文件，替代硬编码列表。
- 其他辅助改进：修复模型检测逻辑、优化 dashboard 生成脚本等，确保 CI 流程顺畅。

评论区精华

由于 review 评论为空，讨论主要体现在提交历史中。例如，在提交 c84f085 中，对 LTX-2 模型的配置进行了性能对比：

> 基准测试显示，torch.compile + ulysses 配置为 32.38 秒，而 torch.compile + CFG parallel 配置为 27.95 秒，因此选择了 CFG parallel 作为优化方案。

这体现了性能调优中的技术权衡。

风险与影响

- 技术风险：端口逻辑变更可能引入新的崩溃点，特别是在严格端口模式下；基准测试配置变更可能影响历史数据比较；自动发现单元测试可能导致未预期文件被包含。
- 影响范围：对终端用户无直接影响；提升 CI 稳定性和基准测试真实性，便于团队检测性能回归；简化测试维护，减少人为错误。

关联脉络

与本 PR 相关的历史 PR 包括：

- PR 22099（对齐扩散模型预设）：同样优化扩散模型基准测试配置。
- PR 22091（扩散模型 NVFP4 默认后端）：涉及扩散模型性能基准测试改进。
- PR 21828（验证注意力后端）：关注扩散模型的后端验证，与本 PR 的测试稳定性增强相辅相成。

这些 PR 共同推动了扩散模型 CI 和性能测试的持续优化。