

PR #29331 完整报告

sgl-project/sglang

[NPU] Add new diffusion tests

合并时间: 2026-07-07 05:40

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29331>

执行摘要

- 一句话: 为 NPU 扩散模型新增 7 个测试用例与性能基线
- 推荐动作: 建议合并此 PR, 但后续应关注 perf_baselines_npu.json 的基线稳定性, 并计划恢复 COSMOS3_NANO 的 NPU 测试。重构后的采样参数提取模式值得推广到其他共享配置中。

功能与动机

PR 作者在描述中指出『需要扩展测试覆盖』(原文: Need to extend test coverage.), 旨在提升 NPU 平台扩散模型测试的覆盖范围, 确保新增模型在 NPU 上的基本功能与性能可被 CI 持续验证。

实现拆解

1. 引入 use_modelscope 辅助函数并新增测试用例: 在 testcase_configs_npu.py 中将原有手动拼接路径统一为 use_modelscope 调用, 降低路径定义冗余; 随后在 ONE_NPU_CASES 和 TWO_NPU_CASES 列表中新增 ERNIE-Image、GLM-Image、FLUX.2-klein-4B、Z-Image、JoyAI-Image-Edit、LTX-2、MOVA-360p 的测试用例, 覆盖 T2I、TI2I、TI2V 等多种任务类型。
2. 提取公共采样参数: 在 testcase_configs.py 中新增 COSMOS3_NANO_CI_sampling_params 常量, 将原来内联在 GPU 测试用例中的参数提取出来, 便于多平台复用。
3. 重构 GPU 测试配置: 在 gpu_cases.py 中引用新常量替换内联定义, 去除重复。
4. 更新 NPU 性能基线: 在 perf_baselines_npu.json 中为新增的 ERNIE-Image、GLM-Image、Z-Image 等模型添加端到端耗时、去噪步耗时等基线数据, 作为 CI 性能回归检测的依据。
5. 更新 GT commit 并移除失败用例: 在 test_utils.py 中更新 NPU 使用的测试数据基线提交 ID; 根据提交历史, 由于运行时失败, 移除了 COSMOS3_NANO 的 NPU 测试。

关键文件:

- python/sglang/multimodal_gen/test/server/ascend/testcase_configs_npu.py (模块 NPU 测试配置; 类别 test; 类型 test-coverage; 符号 use_modelscope): 主要修改: 引入 use_modelscope 辅助函数, 新增 7 个 NPU 扩散测试用例 (ERNIE-Image、GLM-Image、FLUX.2-klein-4B、Z-Image、JoyAI-Image-Edit、LTX-2、MOVA-360p)

- python/sglang/multimodal_gen/test/server/ascend/perf_baselines_npu.json (模块 性能基线; 类别 test; 类型 test-coverage) : 为新增的 NPU 扩散测试添加性能基线数据 (ernie_image_t2i_1npu、glm_image_t2i_1npu、z_image_t2i_1npu 等), 并调整 MOVA 用例名称
- python/sglang/multimodal_gen/test/server/testcase_configs.py (模块 测试配置; 类别 test; 类型 test-coverage) : 新增 COSMOS3_NANO_CI_sampling_params 公共常量, 供 GPU 与 NPU 测试共享
- python/sglang/multimodal_gen/test/server/gpu_cases.py (模块 GPU 测试; 类别 test; 类型 test-coverage) : 引用新常量 COSMOS3_NANO_CI_sampling_params 替换内联参数, 简化代码
- python/sglang/multimodal_gen/test/test_utils.py (模块 测试工具; 类别 test; 类型 test-coverage) : 更新 NPU 测试数据基线 commit ID, 确保测试数据与新增模型匹配

关键符号: use_modelscope

关键源码片段

python/sglang/multimodal_gen/test/server/ascend/testcase_configs_npu.py

主要修改: 引入 use_modelscope 辅助函数, 新增 7 个 NPU 扩散测试用例 (ERNIE-Image、GLM-Image、FLUX.2-klein-4B、Z-Image、JoyAI-Image-Edit、LTX-2、MOVA-360p)

```
import os
from sglang.multimodal_gen.test.server.testcase_configs import (
    DiffusionSamplingParams,
    DiffusionServerArgs,
    DiffusionTestCase,
    T2I_sampling_params,
)

# Modelscope 模型缓存根目录
MODELSCOPE_MODEL_WEIGHTS_DIR = "/root/.cache/modelscope/hub/models/"

def use_modelscope(name: str):
    """返回 modelscope 模型权重完整路径"""
    return os.path.join(MODELSCOPE_MODEL_WEIGHTS_DIR, name)

# 定义所有模型权重路径, 统一通过 use_modelscope 生成
ERNIE_IMAGE_WEIGHTS_PATH = use_modelscope("PaddlePaddle/ERNIE-Image")
FLUX_1_DEV_WEIGHTS_PATH = use_modelscope("black-forest-labs/FLUX.1-dev")
# ... 其他路径定义 ...

EXTRAS_DISABLE_WARMUP = ["--warmup-mode", "request"]

# 单 NPU 测试用例
ONE_NPU_CASES: list[DiffusionTestCase] = [
    # ERNIE-Image T2I 测试, 关闭一致性校验
    DiffusionTestCase(
```

```

    "ernie_image_t2i_1np",
    DiffusionServerArgs(model_path=ERNIE_IMAGE_WEIGHTS_PATH,
                        extras=EXTRAS_DISABLE_WARMUP),
    T2I_sampling_params,
    run_consistency_check=False,
),
# FLUX.1-dev T2I 测试（保留旧用例）
DiffusionTestCase(
    "flux_image_t2i_1np",
    DiffusionServerArgs(model_path=FLUX_1_DEV_WEIGHTS_PATH,
                        extras=EXTRAS_DISABLE_WARMUP),
    T2I_sampling_params,
),
# ... 其他新用例 ...
]

```

评论区精华

PR 提交后 CI 在 [multimodal-gen-test-1-b200](#) 任务中失败，但作者和审核者 [ping1jing2](#) 确认失败与 PR 无关（PR body 中说明『..... it isn't connected with this PR.』），审核者通过 [/tag-and-rerun-ci](#) 触发重试后批准合并。

- CI failure unrelated to this PR (other): 审核者批准合并，未要求额外修改。

风险与影响

- 风险：
 - 性能基线可靠性：perf_baselines_npu.json 中的基线数据来自单次运行，若 NPU 环境存在波动可能导致后续 CI 误报性能回归。建议在稳定环境采集多次数据取中位数。
 - 覆盖盲区：因运行时失败移除了 COSMOS3_NANO 的 NPU 测试，该模型在 NPU 上的覆盖率暂时缺失。需在模型适配问题解决后重新添加。
 - GT commit 更新：test_utils.py 中 GT commit 变更影响所有 NPU 扩散测试的数据一致性，需确保新 commit 包含正确的预期输出。
- 影响：
 - 用户 / 系统：无用户可见变更，仅影响 CI 测试流水线。NPU 扩散模型的回归检测能力显著增强。
 - 团队：为 NPU 团队提供了 7 个新增模型的自动化验证，减少手动测试成本。重构后的 COSMOS3_NANO_CI_sampling_params 可在 GPU 和 NPU 之间共享，提升配置一致性。
 - 影响程度：中等正向影响，测试覆盖扩大有助于提前发现回归缺陷。
 - 风险标记：基线数据依赖单次运行，COSMOS3_NANO NPU 测试移除，GT commit 变更影响数据一致性

关联脉络

- PR #29403 feat: sync npu nightly test improvements from Ascend testcases: 同为 NPU 测试套件增强，当前 PR 在此基础上进一步增加了扩散模型测试覆盖。