PR #25074 完整报告

sgl-project/sglang

[MUSA][22/N] ci(musa): repack wheels with +musa metadata, refine path filters, sync multimodal tests, and add nightly workflow

合并时间：2026-05-22 16:24

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25074

执行摘要

- 一句话：完善 MUSA CI：重构 wheels 元数据、收紧路径过滤、同步测试并添加 nightly 工作流
- 推荐动作：该 PR 是 CI 基础设施的重要改进，值得 SGLang 其他硬件后端（如 NPU、Intel GPU）参考学习。建议审核者关注路径过滤规则的完备性和 wheel 元数据在不同 Python 打包工具下的兼容性。

功能与动机

来自 PR body：'Reduce unnecessary MUSA PR CI runs by tightening the path matching rules to trigger only for MUSA-related edits, and fail fast if dependency installation hangs.' 同时需要保持 MUSA 多模态测试与当前测试布局一致，并增加 nightly 工作流来捕获回归问题。

实现拆解

1. Wheel 元数据重写：修改 rename_wheels_musa.sh，将原始 wheel 解包，在 METADATA 的 Version 字段追加 +musa<数字> 后缀，更新 WHEEL Tag，重命名 .dist-info 目录，然后重新打包并自动重命名输出文件。
2. 收紧 PR 触发路径：在 pr-test-musa.yml 中细化 main_package、multimodal_gen 和 sgl_kernel 的 glob 规则，仅当修改 MUSA 相关文件时才触发 CI；同时为依赖安装步骤添加 10 分钟超时，防止挂起。
3. 同步多模态测试：将原来的 test_server_a_musa.py 和 test_server_b_musa.py 重命名为 test_server_1_gpu_musa.py 和 test_server_1_gpu_musa_nightly.py，统一测试入口；新增 testcase_configs_musa.py 中的 hf_cached_model 函数统一管理模型路径，并分离出 nightly-only 的测试用例集。
4. 新增 nightly 工作流：创建 nightly-test-musa.yml，包含 kernel 测试、multimodal 层测试和 1/2 GPU 服务器测试，通过 cron 定时（UTC 16:00）触发，支持 workflow_dispatch 过滤指定 job，与 PR 工作流独立，降低 PR CI 负担。

关键文件：

- python/sglang/multimodal_gen/test/server/musa/testcase_configs_musa.py（模块 测试配置；类别 test；类型 test-coverage；符号 hf_cached_model）：核心测试配置文件，统

一 MUSA 测试用例定义，引入 hf_cached_model 函数和夜间测试用例集，重构了导入关系。

- .github/workflows/nightly-test-musa.yml（模块 CI 工作流；类别 infra；类型 infrastructure）：新增的 MUSA nightly 工作流，分离 kernel、多模态层和服务器测试，减少 PR CI 负担。
- scripts/ci/musa/rename_wheels_musa.sh（模块 Wheel 打包；类别 infra；类型 infrastructure）：重写 MUSA wheel 打包脚本，使 METADATA 版本字段附带 +musa 后缀，并重命名 dist-info 目录，确保安装正确。
- .github/workflows/pr-test-musa.yml（模块 CI 工作流；类别 infra；类型 infrastructure）：修改 PR 触发路径过滤规则，增加 MUSA 后端路径、收紧过滤规则，减少不必要的 CI 运行。

关键符号：hf_cached_model, TestDiffusionServerOneGpuMusaNightly, TestDiffusionServerOneGpuMusa, TestDiffusionServerTwoGpuMusa

关键源码片段

python/sglang/multimodal_gen/test/server/musa/testcase_configs_musa.py

核心测试配置文件，统一 MUSA 测试用例定义，引入 hf_cached_model 函数和夜间测试用例集，重构了导入关系。

```
# 引入 functools.lru_cache 来缓存 HuggingFace 模型快照路径
from functools import lru_cache

@lru_cache(maxsize=None)
def hf_cached_model(repo_id: str) -> str:
    """将一个 HF repo id 解析为本地缓存快照路径，用于 MUSA 运行器。"""
    # 延迟导入 huggingface_hub 以避免不必要的开销
    from huggingface_hub import snapshot_download
    # local_files_only=True 确保不会因网络问题挂起，因为 CI 环境已离线缓存好模型
    return snapshot_download(repo_id, local_files_only=True)

# 配置 MUSA 专用的 TI2I 采样参数，使用固定测试图片路径
MUSA_TI2I_sampling_params = replace(
    TI2I_sampling_params,
    image_path="/hf-cache/hub/musa-test-assets/TI2I_Qwen_Image_Edit_Input.jpg",
)

# 单 GPU 测试用例列表：PR CI 运行时仅执行这些用例
ONE_GPU_MUSA_CASES: list[DiffusionTestCase] = [
    DiffusionTestCase(
        "qwen_image_t2i_musa",
        DiffusionServerArgs(
            model_path=hf_cached_model("Qwen/Qwen-Image"), # 通过缓存函数获取本地路径
            modality="image",
        ),
        T2I_sampling_params,
        run_consistency_check=False,
```

```
    ),
    DiffusionTestCase(
        "wan2_1_t2v_1.3b_musa",
        DiffusionServerArgs(
            model_path=hf_cached_model("Wan-AI/Wan2.1-T2V-1.3B-Diffusers"),
            modality="video",
            custom_validator="video",
        ),
        DiffusionSamplingParams(prompt=T2V_PROMPT),
        run_consistency_check=False,
    ),
]
```

评论区精华

Review 中主要讨论了以下方面：

- 模型路径管理：yeahdongcn 建议使用环境变量 HF_HUB_OFFLINE=1 和 HF_HOME=/hf-cache 代替硬编码路径，开发者采纳。
- 路径过滤排序：yeahdongcn 建议按字母顺序排列触发路径以方便维护，后续提交已调整。
- 超时设置：对 PR 工作流中依赖安装的 10 分钟超时是否充足提出疑问，未明确结论；nightly 工作流中某些步骤原本缺少超时，已补上。
- Wheel 脚本来源：yeahdongcn 指出重命名脚本改编自 PR #25573，并表达感谢，无争议。
  - 模型路径从硬编码改为环境变量驱动 (design): 开发者采纳，在测试配置和环境变量中统一使用 HF_HOME 和 HF_HUB_OFFLINE。
  - 安装依赖超时 10 分钟是否足够 (question): 未在评论中看到进一步答复，超时设置保留为 10 分钟。
  - 按字母顺序排列触发路径 (style): 采纳，后续提交中路径已按字母顺序排列。
  - Nightly 工作流步骤缺少超时 (other): 开发者添加了超时设置。
  - Wheel 重命名脚本来源 (other): 承认来源，无争议。

风险与影响

- 风险：
  1. Wheel 兼容性：修改 METADATA 和 dist-info 目录名可能导致部分安装工具或依赖解析异常，若 +musa 后缀不被某些工具识别，可能影响用户安装。
  2. 路径过滤遗漏：新的 glob 规则可能因模式不全面而遗漏 MUSA 相关变更，导致应该运行的 CI 未触发。
  3. Nightly 资源浪费：若 nightly 工作流中测试用例过多或超时设置过长，可能持续占用 CI 资源而无人监控结果。
  4. 测试用例合并遗漏：将多组测试用例合并进同一文件（如 ONE_GPU_MUSA_CASES）时，若参数配置有误，可能导致原本覆盖的测试场景丢失。 - 影响：范围：仅影响 MUSA 硬件平台的 CI 流程和 wheel 构建，不涉及核心推理逻辑。影响：

- 对开发者：MUSA PR CI 触发更精准，减少等待时间；nightly 测试覆盖更大，有助于提前发现回归。
- 对 MUSA 用户：wheel 命名和元数据更规范，确保安装正确。
- 对 CI 运维：新增一个 nightly 工作流，需关注资源消耗和结果通知。
- 风险标记：Wheel 元数据兼容性，路径过滤遗漏，Nightly 资源消耗，超时限制不统一

关联脉络

- PR #25573 MUSA wheel rename and build: wheel 重命名脚本改编自此 PR