PR #22560 完整报告

sgl-project/sglang

[Diffusion][CI] Fix nunchaku unit test broken by #22365

合并时间：2026-04-11 08:49

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22560

执行摘要

本 PR 修复了扩散模型量化单元测试因 PR #22365 引入的 CI 失败，通过 Mock maybe_download_model 函数避免测试中使用的本地临时路径被误识别为 Hugging Face 仓库而尝试下载。变更仅涉及一个测试文件，风险极低，旨在恢复 CI 测试稳定性。

功能与动机

修复 multimodal-gen-unit-test CI 任务的失败。PR body 指出：PR #22365 新增的 _resolve_quant_config_from_transformer_override 函数会调用 maybe_download_model 处理 transformer 权重路径，但单元测试 test_resolve_transformer_quant_load_spec_keeps_nunchaku_hook 中使用的路径 /tmp/svdq-int4_r32.safetensors 是一个本地假路径，不存在于 HF 仓库，导致 HF Hub 验证失败。作者提供了 CI 失败示例链接，并明确修复目标为使该测试任务通过。

实现拆解

仅修改文件 python/sglang/multimodal_gen/test/unit/test_transformer_quant.py，在测试函数 test_resolve_transformer_quant_load_spec_keeps_nunchaku_hook 上添加一个 Mock 装饰器：

```
@patch(
    "sglang.multimodal_gen.runtime.loader.transformer_load_utils.maybe_download_model",
    side_effect=lambda path, **kw: path,
)
```

该 Mock 将 maybe_download_model 的行为替换为直接返回输入路径，避免实际下载操作，从而绕过 HF 验证失败问题。

评论区精华

Review 中仅有一名审核者（yhyang201）批准，无具体评论。从上下文看，修复策略直接，未引发技术争议。

风险与影响

- 风险：极低。Mock 仅影响测试环境，不改变生产代码；变更范围小，易于验证。潜在风险是 Mock 可能掩盖 maybe_download_model 在实际使用中的问题，但鉴于这是针对测试假路径的修复，风险可控。

- 影响：修复 CI 失败，确保扩散模型量化相关单元测试通过，维护测试套件稳定性。对用户和系统无直接影响。

关联脉络

- 直接关联：PR #22365（标题未知）是本修复的诱因，其引入的 *_resolve_quant_config_from_transformer_override* 函数导致测试失败。
- 仓库趋势：近期多个 PR 涉及 CI 测试修复和扩散模型模块（如 PR #22460、#21960），显示团队在加强测试覆盖和 CI 稳定性，本 PR 是这一趋势中的一个小幅维护性修复。