

PR #51427 完整报告

vllm-project/vllm

[Bugfix][models_multimodal] Remote HF python code misses importing class

合并时间: 2026-08-08 12:09

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/51427>

执行摘要

- 一句话: 修复 MiniCPM-V 远程代码 List 未导入导致的 NameError
- 推荐动作: 这是一个小而清晰的临时补丁 PR, 通读一遍即可, 无需精读。值得关注的设计点: 在 conftest 层用 hasattr 守卫注入缺失别名, 兼顾幂等性与最小侵入; 明确的 TODO 与回退约定; 与上游 HF 修复 PR 的联动跟踪。若你负责多模态测试或维护第三方远程代码兼容性, 可借鉴这种“测试环境自愈”思路, 同时注意补丁的全局命名空间影响。

功能与动机

PR #48413 合并后在 Multi-Modal Models (Extended Generation 3) 测试组引入两个独立运行时错误, 本 PR 处理第一个: NameError: name 'List' is not defined。根因是远程 HF 文件 resampler.py 定义函数时返回类型标注为 List, 却没有 import 该类; 旧版本 torch 会从 torch.nn.functional 的星号导入中泄漏 List 别名, 新版 torch 不再如此。维护者 DarkLight1337 在评论中明确“先修 CI, 等上游修复后回退”, 因此需要本地临时补丁保持 CI 绿色。

实现拆解

本 PR 只改动 1 个文件、净增 9 行, 属于测试脚手架层面的兼容性补丁, 实现路径如下:

1. 定位问题: 合并 #48413 后, models/multimodal/generation/test_common.py 中 minicpmv_25 的 3 个用例在加载 MiniCPM-V 2.5 远程代码时全部报 NameError: name 'List' is not defined, CI 组为 Multi-Modal Models (Extended Generation 3)。
2. 根因确认: openbmb/MiniCPM-Llama3-V-2_5 的远程 resampler.py 用 List 做返回类型标注但模块内未导入; 此前该名字从 from torch.nn.functional import * 泄漏出来, 新版 torch 改用 builtin generics 后不再复导出 typing 别名。
3. 补丁实现: 在 tests/models/multimodal/conftest.py 顶部导入 torch.nn.functional as F, 并在模块加载时执行 if not hasattr(F, "List"): F.List = list, 以属性注入方式恢复兼容别名; hasattr 守卫保证幂等, 不会覆盖 torch 自身的定义。补丁发生在 pytest collection 之前, 确保后续所有多模态测试共享该兼容环境。
4. 验证: 作者在本地对 test_common.py 的 test_case100/101/102 三个用例做 before/after 对比 (失败转通过), 并由 AndreasKaratzas 触发 Buildkite CI #82897 验证全量信号。

5. 回退策略：补丁以 TODO 注释标记，等待 HF 侧 checkpoint 修复 resampler.py 的 import 后移除；同时 PR #48413 引入的第二个错误（MiniCPMV.all_tied_weights_keys AttributeError）不在本 PR 范围，由后续 PR 单独处理。

关键文件：

- tests/models/multimodal/conftest.py（模块测试配置；类别 test；类型 test-coverage）：唯一变更文件。在 pytest 配置模块顶层为 torch.nn.functional 补回 List 别名，修复 MiniCPM-V 2.5 远程代码加载时的 NameError，是保持 CI 绿色的关键补丁点。

关键符号：未识别

关键源码片段

tests/models/multimodal/conftest.py

唯一变更文件。在 pytest 配置模块顶层为 torch.nn.functional 补回 List 别名，修复 MiniCPM-V 2.5 远程代码加载时的 NameError，是保持 CI 绿色的关键补丁点。

```
# tests/models/multimodal/conftest.py（节选）
# SPDX-License-Identifier: Apache-2.0
# SPDX-FileCopyrightText: Copyright contributors to the vLLM project
"""Pytest configuration for vLLM multimodal tests."""

import os
import warnings

import torch
import torch.nn.functional as F

from vllm.platforms import current_platform

# openbmb/MiniCPM-Llama3-V-2_5 的远程 resampler.py 用 List 做类型标注
# 却没有 import，此前依赖 ``from torch.nn.functional import *`` 泄漏出的别名。
# 新版 torch 改用 builtin generics，不再复导出 typing 别名，
# 因此执行远程代码时会抛出 NameError。这里手动恢复兼容别名，
# 仅在属性缺失时注入，避免覆盖 torch 自身的定义。
# TODO: 上游 checkpoint 修复 import 后移除本补丁。
if not hasattr(F, "List"):
    F.List = list

def pytest_configure(config):
    """Early ROCm configuration that must happen before test collection."""
    if not current_platform.is_rocm():
        return

    # 关掉 ROCm 上的 skinny GEMM，避免 wvSplitKrc 原子归约导致结果不确定
    # 参见 vllm PR #33493 的讨论
    os.environ["VLLM_ROCM_USE_SKINNY_GEMM"] = "0"
    warnings.warn(
```

```
"ROCM: Set VLLM_ROCM_USE_SKINNY_GEMM=0 to avoid non-deterministic "  
"results from skinny GEMM atomic reductions",  
UserWarning,  
stacklevel=1,  
)
```

评论区精华

核心讨论围绕“等待上游修复还是本地打补丁”展开：

- DarkLight1337 首先询问维护者：“@tc-mb 你会在 HF Hub 上很快修复吗？如果不会，我们可能需要 patch 或禁用测试来保持 CI 绿色。”
- gchinora 回应 HF 侧已有修复 PR: https://huggingface.co/openbmb/MiniCPM-Llama3-V-2_5/discussions/79/files
- DarkLight1337 最终拍板：“先修 CI，等上游修复后再回退”，即接受本地临时补丁方案。
- 上游修复 vs 本地补丁的方案抉择 (design): 采用 conftest 中注入 F.List 的临时补丁，并标记 TODO 等待上游修复后 revert。
- CI 验证与回退时机 (testing): CI 通过后合入补丁，回退时机以上游 HF checkpoint 修复为准。

风险与影响

- 风险：风险点如下，整体可控：
- monkey-patch 全局性：给 `torch.nn.functional` 注入 `List` 属性会影响测试进程内所有模块的命名空间，但 `hasattr` 守卫 + `list` 赋值使其幂等且低风险；不过补丁位于 `conftest` 模块顶层，任何依赖 `F.List` 缺失行为的测试都可能被掩盖。
- 版本兼容性：若未来 torch 重新复导出 `List` 或改变模块结构，`hasattr` 判断能保证逻辑安全；但若 torch 内部将 `F.List` 用于其他用途，注入可能产生意外交互（当前未见此类用法）。
- 遗留问题：PR #48413 引入的 `MiniCPMV.all_tied_weights_keys` `AttributeError` 未在本 PR 修复，`MiniCPM-V 2.5` 相关测试组仍可能继续失败，需配套 PR 跟进。
- 回退依赖：补丁的永久化风险——若上游 HF checkpoint 修复迟迟不合入，或合入后团队忘记 revert，该兼容层会长期残留并掩盖同类问题。建议在 TODO 中挂接上游 PR 跟踪。
 - 不涉及生产推理路径，无性能、安全影响。
- 影响：影响范围集中在测试与 CI 链路：
- 对用户：无直接功能影响，补丁不触达任何推理或模型加载代码。
- 对 CI/ 测试：恢复 AMD CI 的 Multi-Modal Extended Generation 3 测试组绿色，消除 `minicpmv_25` 三个用例的确定性失败；Buildkite CI #82897 已验证。
- 对团队：需要跟踪 HF 上游修复进度并及时 revert；为后续同类“远程代码不可控”场景提供了一个可复用的测试脚手架补丁范式。
- 风险标记：临时补丁需上游修复后回退，monkey-patch 影响测试进程全局命名空间，遗留 `MiniCPMV` `AttributeError` 未修

关联脉络

- PR #48413（标题未在上下文中提供，见 PR body 引用）：PR body 明确指出本 PR 修复了 #48413 合并后引入的 `NameError`；#48413 将 MiniCPM-V 2.5 带入 Multi-Modal Extended Gen 3 测试组，并遗留 `all_tied_weights_keys` `AttributeError` 待后续 PR 修复，是本 PR 的直接关联源头。