

# PR #42040 完整报告

vllm-project/vllm

[Bugfix] Fix LOGITPROC\_SOURCE\_ENTRYPOINT test to use spawn-compatible dist-info registration for XPU/ROCm

合并时间: 2026-05-09 12:32

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/42040>

## 执行摘要

- 一句话: 使用 dist-info 注册修复 XPU/ROCm 上 logitproc 测试的 spawn 兼容性
- 推荐动作: 本 PR 虽然只涉及测试代码, 但采用的通过磁盘 dist-info 包模拟 entrypoint 注册的方案具有通用性, 未来需要实现多进程环境下的 entrypoint 测试时可借鉴。同时, cross-platform 的 fork/spawn 策略设计清晰, 值得测试架构开发者精读。

## 功能与动机

在 XPU/ROCm 平台, 子进程必须使用 spawn 模式而非 fork。原测试依靠 fork 才能将 monkey-patch 的 importlib.metadata.entry\_points 传递给 worker, 导致 test\_custom\_logitsprocs[LOGITPROC\_SOURCE\_ENTRYPOINT] 和 test\_rejects\_custom\_logitsprocs[LOGITPROC\_SOURCE\_ENTRYPOINT] 在这两个平台上持续失败。此 PR 是 #41423 的 follow-up, 旨在解除 XPU/ROCm CI 的测试阻塞。

## 实现拆解

1. 新增 spawn 兼容入口: 在 tests/v1/logits\_processors/utils.py 中实现 register\_fake\_entrypoint(monkeypatch), 创建临时 .dist-info 目录并写入 METADATA 和 entry\_points.txt, 通过 monkeypatch.setenv('PYTHONPATH', ...) 暴露给子进程, 同时用 monkeypatch.syspath\_prepend 更新当前进程的 sys.path。
2. 封装平台感知函数: 在同一个文件中添加 setup\_fake\_entrypoint(monkeypatch), 它根据 requires\_spawn\_multiprocessing() (定义于 tests/utils.py) 自动选择: 若平台需要 spawn, 则调用 register\_fake\_entrypoint 并设置 VLLM\_WORKER\_MULTIPROC\_METHOD=spawn; 否则沿用原有 monkeypatch.setattr(importlib.metadata, 'entry\_points', fake\_entry\_points) 并设置为 fork。
3. 更新消费端测试: 修改 tests/v1/logits\_processors/test\_custom\_offline.py 和 test\_custom\_online.py, 将原先直接 import fake\_entry\_points 并在测试函数中手动 patch 的代码统一替换为调用 setup\_fake\_entrypoint(monkeypatch), 并移除了强制 VLLM\_WORKER\_MULTIPROC\_METHOD=fork 的设定。
4. 增强测试框架基础: 在 tests/utils.py 中新增 requires\_spawn\_multiprocessing() 函数, 重构 RemoteOpenAIServerCustom.\_start\_server() 使其根据平台选择 spawn 或 fork, 并引入 \_run\_in\_new\_process\_group 和 \_terminate\_process\_tree 来改善进程组启动与清理

逻辑，确保子进程被正确终止。

5. 保留向后兼容：所有临时目录通过 monkeypatch 管理，测试退出后自动清理；fork 模式下的原有行为完全不变。

关键文件：

- tests/v1/logits\_processors/utils.py（模块 测试工具；类别 test；类型 test-coverage；符号 register\_fake\_entrypoint, fake\_entry\_points, setup\_fake\_entrypoint）：核心变更文件，新增 register\_fake\_entrypoint 和 setup\_fake\_entrypoint 函数，实现 spawn 兼容的 dist-info 注册。
- tests/utils.py（模块 测试工具；类别 test；类型 test-coverage；符号 requires\_spawn\_multiprocessing, \_run\_in\_new\_process\_group, exit, \_terminate\_process\_tree）：提供 requires\_spawn\_multiprocessing 辅助函数，并重构 RemoteOpenAIServerCustom 的子进程启动和销毁逻辑以支持 spawn 模式。
- tests/v1/logits\_processors/test\_custom\_offline.py（模块 离线测试；类别 test；类型 test-coverage）：应用了新的 setup\_fake\_entrypoint 函数替换原有的 monkey-patch，是变更的消费方之一。
- tests/v1/logits\_processors/test\_custom\_online.py（模块 在线测试；类别 test；类型 test-coverage）：在线测试版本，同样应用 setup\_fake\_entrypoint 替换 monkey-patch。

关键符号：register\_fake\_entrypoint, setup\_fake\_entrypoint, fake\_entry\_points, requires\_spawn\_multiprocessing, \_run\_in\_new\_process\_group, \_terminate\_process\_tree

## 关键源码片段

### tests/v1/logits\_processors/utils.py

核心变更文件，新增 register\_fake\_entrypoint 和 setup\_fake\_entrypoint 函数，实现 spawn 兼容的 dist-info 注册。

```
# tests/v1/logits_processors/utils.py (head 版本核心片段)
import os
import tempfile
from pathlib import Path

def register_fake_entrypoint(monkeypatch) -> str:
    """为 spawn 子进程创建真实的 .dist-info 包，用于模拟 entrypoint。

    该函数创建临时目录并写入 METADATA 和 entry_points.txt，
    然后通过 PYTHONPATH 和 sys.path 暴露给子进程和当前进程。
    """
    tmpdir = Path(tempfile.mkdtemp(prefix="dummy-logitproc-"))
    dist_info = tmpdir / "dummy_logitproc-0.1.dist-info"
    dist_info.mkdir()

    # METADATA 是 importlib.metadata 必需的描述文件
    (dist_info / "METADATA").write_text(
        "Metadata-Version: 2.1\nName: dummy-logitproc\nVersion: 0.1\n",
```

```

        encoding="utf-8",
    )

    # entry_points.txt 声明可用的 entrypoint 组及对应 FQCN
    (dist_info / "entry_points.txt").write_text(
        f"[{LOGITSPROCS_GROUP}]\n"
        f"{DUMMY_LOGITPROC_ENTRYPOINT} = {DUMMY_LOGITPROC_FQCN}\n",
        encoding="utf-8",
    )

    # 将临时目录加入 PYTHONPATH, 使 spawn 子进程能发现此包
    existing = os.environ.get("PYTHONPATH", "")
    monkeypatch.setenv(
        "PYTHONPATH", str(tmpdir) + (os.pathsep + existing if existing else "")
    )

    # 将临时目录前置到 sys.path, 使当前 driver 进程也能发现 entrypoint
    monkeypatch.syspath_prepend(str(tmpdir))

    return str(tmpdir)

def setup_fake_entrypoint(monkeypatch) -> None:
    """根据平台选择合适的 entrypoint 注册方式。

    若平台需要 spawn (如 XPU/ROCm), 则创建真实 dist-info 包;
    否则沿用传统的 monkey-patch 方式, 仅对 fork 有效。
    """
    if requires_spawn_multiprocessing():
        register_fake_entrypoint(monkeypatch)
        monkeypatch.setenv("VLLM_WORKER_MULTIPROC_METHOD", "spawn")
        return

    # 传统 fork 模式: 直接替换 importlib.metadata.entry_points
    import importlib.metadata
    monkeypatch.setattr(importlib.metadata, "entry_points", fake_entry_points)
    monkeypatch.setenv("VLLM_WORKER_MULTIPROC_METHOD", "fork")

```

## tests/utils.py

提供 `requires_spawn_multiprocessing` 辅助函数, 并重构 `RemoteOpenAIServerCustom` 的子进程启动和销毁逻辑以支持 `spawn` 模式。

```

# tests/utils.py (head 版本关键片段)
def requires_spawn_multiprocessing() -> bool:
    """判断当前平台是否需要使用 spawn 模式替代 fork。"""
    return current_platform.is_rocm() or current_platform.is_xpu()

def _run_in_new_process_group(

```

```

child_process_fxn: Callable[[dict[str, str] | None, str, list[str]], None],
env_dict: dict[str, str] | None,
model: str,
vllm_serve_args: list[str],
) -> None:
    """在新进程组中执行子进程函数，便于后续终止整组进程。"""
    os.setsid()
    child_process_fxn(env_dict, model, vllm_serve_args)

class RemoteOpenAIServerCustom(RemoteOpenAIServer):
    # ... 其他方法 ...

    def _start_server(
        self, model: str, vllm_serve_args: list[str], env_dict: dict[str, str] | None
    ) -> None:
        # 根据平台选择进程启动方式
        method = "spawn" if requires_spawn_multiprocessing() else "fork"
        ctx = get_context(method)
        self.proc: Process = cast(Any, ctx).Process(
            target=_run_in_new_process_group,
            args=(self.child_process_fxn, env_dict, model, vllm_serve_args),
        )
        self.proc.start()

    def _terminate_process_tree(self) -> None:
        """终止整个进程组，确保清理干净。"""
        pid = self.proc.pid
        if pid is None:
            return
        pgid: int | None
        try:
            pgid = os.getpgid(pid)
            if pgid != pid:
                pgid = None
        except (ProcessLookupError, OSError):
            pgid = None
        with contextlib.suppress(ProcessLookupError, OSError):
            self.proc.terminate()
        self.proc.join(15)
        if self.proc.is_alive():
            if pgid is not None:
                with contextlib.suppress(ProcessLookupError, OSError):
                    os.killpg(pgid, signal.SIGKILL)
            else:
                self.proc.kill()
        self.proc.join(10)
        self._kill_process_group_survivors(pgid)

```

## 评论区精华

- gemini-code-assist[bot] 指出 driver 进程也需要发现 entrypoint (地点: tests/v1/logits\_processors/utils.py line ~219) : “While updating PYTHONPATH ensures that spawned subprocesses can discover the fake entry point, the current process’s sys.path remains unchanged… you should also update sys.path for the current process.”该建议被采纳, 后续 commit 添加了 monkeypatch.syspath\_prepend(str(tmpdir))。
- AndreasKaratzas 关于 CI 阻塞的协调: 由于 AMD CI 长时间阻塞, Andreas 另开了 #42115 包含同样修复, 但后因作者回归而关闭, 未造成冲突。
  - driver 进程也需要发现 entrypoint, 需更新 sys.path (correctness): 已采纳, 增加了 monkeypatch.syspath\_prepend(str(tmpdir)) 调用。
  - CI 阻塞导致另开重复 PR (other): 作者 rebase 后 CI 通过, PR #42115 关闭。
  - XPU 上 LoRA 测试失败是否可忽略 (other): rebase 后测试通过, LoRA 失败已由其他 PR 处理。

## 风险与影响

- 风险: 主要风险在于临时 dist-info 目录可能未被正确清理, 但 monkeypatch 机制保证了作用域内清理。另外, 修改 RemoteOpenAIServerCustom 的子进程启动方式可能影响其他依赖于该类的测试 (如在线 server 测试), 但变更只增强了进程组终止逻辑, 且原有接口不变, 风险较低。新增的 \_terminate\_process\_tree 通过进程组发送 SIGKILL, 在极端情况下可能误杀 pytest 自身的进程组, 但函数中已通过 if pgid != pid 检查来避免。
- 影响: 影响范围限于测试基础设施, 具体为自定义 logits processor 相关的离线与在线测试在 XPU/ROCm 平台上的运行。对其他平台和功能无直接影响。对团队而言, 解除了 XPU/ROCm CI 中的测试阻塞, 合并后可恢复这部分 CI 信号。也提供了一个可复用的模式: 以后任何需要多进程 entrypoint 测试的场景都可借鉴 register\_fake\_entrypoint。
- 风险标记: 测试文件变更, 多进程兼容性, 临时文件管理

## 关联脉络

- PR #41423 Earlier PR introducing custom logits processor entrypoint tests: 此 PR 作为 #41423 的后续, 修复了其引入的 entrypoint 测试在 spawn 模式下的兼容性问题。
- PR #41895 Issue discussing the test failures on XPU/ROCm: 该 issue 讨论了 XPU/ROCm 上 entrypoint 测试失败的问题并催生了本 PR。
- PR #42115 Duplicate PR with same fix: 作者因 CI 阻塞另开的 PR, 与本 PR 内容基本相同。