# PR #44345 完整报告

vllm-project/vllm

[BugFix] Fix sparse NCCL weight transfer test construction

合并时间：2026-06-03 05:51

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44345

## 执行摘要

- 一句话：修复稀疏 NCCL 权重传输测试构造
- 推荐动作：本 PR 为常规 bugfix，变更简单直接，值得快速合并以恢复 CI 稳定性。可关注后续对 NCCLWeightTransferEngine 构造签名的进一步演进。

## 功能与动机

修复 #44272 中测试遗漏：生产路径已通过 WeightTransferEngineFactory.create_engine(...) 传入 model，但测试和文档示例仍直接构造 NCCLWeightTransferEngine(config, parallel_config)，导致 nightly CI 失败。

## 实现拆解

1. 测试修复：在 tests/distributed/test_weight_transfer.py 中，为 test_nccl_receive_sparse_weights_without_init_raises 和 inference_receive_sparse_tensor 两处直接构造 NCCLWeightTransferEngine 的调用补传 MagicMock(spec=torch.nn.Module) 作为第三个参数。
2. 文档同步：在 docs/training/weight_transfer/base.md 的示例代码中，将 WeightTransferEngineFactory.create_engine(...) 调用补上 model=model 参数。

关键文件：

- tests/distributed/test_weight_transfer.py（模块 权重传输；类别 test；类型 test-coverage）：两处直接构造 NCCLWeightTransferEngine 的调用遗漏了新增的 model 参数，导致 nightly CI 失败。通过补传 MagicMock 修复。
- docs/training/weight_transfer/base.md（模块 文档；类别 docs；类型 documentation）：文档示例代码中 WeightTransferEngineFactory.create_engine 调用缺少 model 参数，同步补全以匹配新的构造签名。

关键符号：未识别

## 关键源码片段

### tests/distributed/test_weight_transfer.py

两处直接构造 NCCLWeightTransferEngine 的调用遗漏了新增的 model 参数，导致 nightly CI 失败。通过补传 MagicMock 修复。

```
# tests/distributed/test_weight_transfer.py
# 修复前：engine = NCCLWeightTransferEngine(config, parallel_config)
# 修复后：传入 MagicMock 作为 model 参数

# 测试 1: test_nccl_receive_sparse_weights_without_init_raises
engine = NCCLWeightTransferEngine(
    config, parallel_config, MagicMock(spec=torch.nn.Module)
)
# 后续逻辑不变 ...

# 测试 2: inference_receive_sparse_tensor
engine = NCCLWeightTransferEngine(
    config, parallel_config, MagicMock(spec=torch.nn.Module)
)
```

## 评论区精华

该 PR 无 review 评论，两位 reviewer 直接批准。

- 暂无高价值评论线程

## 风险与影响

- 风险：变更仅涉及测试代码和文档示例，不触及生产逻辑，回归风险极低。稀疏测试仅在 GPU 数量 > 2 时运行（如 PR body 所述），可能无法被单 GPU CI 覆盖，但已通过两 GPU 环境验证。
- 影响：仅影响测试通过率和文档正确性，对用户和系统无直接功能影响。修复后 nightly CI 中稀疏 NCCL 测试可正常通过。
- 风险标记：测试仅在多 GPU 环境运行

## 关联脉络

- PR #44272 [PR #44272 标题未知 ]: 本 PR 修复了该 PR 引入的测试构造遗漏，属于配套修复。