PR #46771 完整报告

vllm-project/vllm

[ModelRunner V2] Update scheduler tests to cover MRV2 paths

合并时间：2026-06-27 00:36

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/46771

执行摘要

- 一句话：调度器测试覆盖 MRV2 路径
- 推荐动作：建议精读测试中针对 V2 行为差异的断言逻辑，理解 V2 model runner 将抢占恢复视为新请求而非缓存请求的设计意图。

功能与动机

V2 model runner 引入后，现有调度器测试仅覆盖 V1 路径，缺少对 V2 行为（如抢占恢复请求的表示方式）的验证。PR 通过参数化测试填补这一空白，提高测试覆盖率。

实现拆解

1. 在 tests/v1/core/utils.py 的 *create_scheduler* 函数新增 use_v2_model_runner 参数，若未指定则读取 envs.VLLM_USE_V2_MODEL_RUNNER 环境变量，并赋值给 scheduler.use_v2_model_runner。
2. 在 tests/v1/core/test_scheduler.py 的 test_cached_request_data_resumed_all_token_ids_mrv1_only 中显式传入 use_v2_model_runner=False，确保该测试始终运行在 V1 模式下。
3. 在 test_kv_connector_handles_preemption 测试函数上新增 use_v2_model_runner 参数化，并在创建调度器时传入该参数；根据 use_v2_model_runner 调整断言：V2 模式下，恢复的请求应出现在 scheduled_new_reqs 而非 scheduled_cached_reqs。
4. 同步更新 *create_scheduler_with_priority* 助手函数，同样支持 use_v2_model_runner 参数。

关键文件：

- tests/v1/core/test_scheduler.py（模块 调度器；类别 test；类型 test-coverage；符号 test_kv_connector_handles_preemption）：核心测试文件，新增 use_v2_model_runner 参数化覆盖 V2 路径，调整断言以匹配 V2 行为。
- tests/v1/core/utils.py（模块 测试工具；类别 test；类型 test-coverage）：工具函数文件，*create_scheduler* 新增 use_v2_model_runner 参数，支持注入 V2 behavior 以方便测试。

关键符号：*create_scheduler*, test_kv_connector_handles_preemption

关键源码片段

tests/v1/core/test_scheduler.py

核心测试文件，新增 use_v2_model_runner 参数化覆盖 V2 路径，调整断言以匹配 V2 行为。

```
@pytest.mark.parametrize("use_v2_model_runner", [False, True])
@pytest.mark.parametrize("is_async", [False, True])
@pytest.mark.parametrize(
    "use_ec_connector, ec_role", [(False, None), (True, "ec_consumer")]
)
def test_kv_connector_handles_preemption(
    is_async, use_ec_connector, ec_role, use_v2_model_runner
):
    """Test whether scheduler with KVConnector can handle preemption when
    blocks run out in allocate_slots()."""
    scheduler = create_scheduler(
        is_async,
        use_kv_connector=mock_kv(matched_tokens=0, is_async=is_async),
        num_blocks=4,
        block_size=16,
        use_ec_connector=use_ec_connector,
        ec_role=ec_role,
        use_v2_model_runner=use_v2_model_runner,
    )
    # ... (request setup and scheduling steps omitted for brevity)
    # After one request completes and triggers preemption:
    if use_v2_model_runner:
        # V2 model runner 将抢占恢复的请求视为 NewRequestData
        # 而不是 CachedRequestData，因此 expects 中 cached_reqs 为 0
        assert output.scheduled_cached_reqs.num_reqs == 0
        assert len(output.scheduled_new_reqs) == 1
    else:
        # V1 model runner 保持原有行为：恢复的请求作为 CachedRequestData
        assert output.scheduled_cached_reqs.num_reqs == 1
        assert output.scheduled_new_reqs == []
```

tests/v1/core/utils.py

工具函数文件，create_scheduler 新增 use_v2_model_runner 参数，支持注入 V2 behavior 以方便测试。

```
def create_scheduler(
    model: str = "facebook/opt-125m",
    # ... other params ...
    use_v2_model_runner: bool | None = None, # 新增参数：控制 scheduler 使用的 model runner 版本
) -> Scheduler | AsyncScheduler:
    # ... config setup ...
    scheduler = scheduler_cls(
        vllm_config=vllm_config,
        kv_cache_config=kv_cache_config,
        block_size=block_size,
```

```
        log_stats=True,
        structured_output_manager=StructuredOutputManager(vllm_config),
    )
    if use_v2_model_runner is None:
        # 默认从环境变量 VLLM_USE_V2_MODEL_RUNNER 读取
        use_v2_model_runner = bool(envs.VLLM_USE_V2_MODEL_RUNNER)
    scheduler.use_v2_model_runner = use_v2_model_runner
    return scheduler
```

评论区精华

无讨论，PR 由 reviewer WoosukKwon 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅涉及测试代码和辅助函数，不修改任何生产逻辑。参数化后不会影响现有 V1 测试结果，新增 V2 路径验证可提前发现问题。
- 影响：影响范围限于 v1 scheduler 的测试覆盖。增强了对 V2 model runner 行为的回归保护，尤其是抢占恢复场景，有助于防止未来重构引入的兼容性问题。
- 风险标记：暂无

关联脉络

- PR #46773 [ModelRunner V2] Fix whisper test: 同属 MRV2 测试修复系列，涉及 V2 model runner 的测试补强。