

PR #57044 完整报告

vllm-project/vllm

[CI] Keep Qwen3 Omni DSpark config fixture complete

合并时间: 2026-09-16 02:19

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/57044>

执行摘要

- 一句话: 用真实配置类修复 Qwen3 Omni DSpark 测试夹具
- 推荐动作: 值得快速浏览而非精读: 核心价值不在 3 行 diff, 而是一个可复用的测试维护模式——mock 只覆盖被测对象, 生产配置契约用真实默认对象承担, 避免 mock 漂移。若想深入理解, 应结合 #56387 的 `_get_dspark_parallel_config` 与 EPLB 状态初始化逻辑一起阅读, 本 PR 是它的测试侧补全。

功能与动机

PR body 明确指出: #56387 合并后, DSpark loader 从 target 的 EPLB 字段推导 draft 配置, H200 Model Executor lane 在构建 #89112 和 #89113 中确定性失败, 因为测试 fixture 只提供了 `tensor_parallel_size`, 而 `_get_dspark_parallel_config` 现在还会读取 `enable_eplb`、`eplb_config`、`enable_elastic_ep`。作者还强调该 fixture 此前已在 #56011、#56600 之后做过单字段修补, 因此改用真实默认 `ParallelConfig` 让生产配置契约完整, 避免下一次新增并行字段再次造成 stale-fixture 失败。

实现拆解

1. 定位根因: PR #56387 修改了 `vllm/v1/worker/gpu/spec_decode/dspark/utils.py`, `_get_dspark_parallel_config` 从 target 的 `parallel_config` 读取 `enable_eplb`、`eplb_config`、`enable_elastic_ep` 三个 EPLB 字段; 而 `tests/model_executor/test_qwen3_omni.py` 中 `test_dspark_shares_target_embedding_with_smaller_draft_vocabulary` 用 `SimpleNamespace(tensor_parallel_size=1)` 构造 `parallel_config`, 缺字段导致属性访问失败, H200 与 MI300 Model Executor CI lane 确定性失败 (构建 #89112、#89113)。
2. 最小修复: 在测试文件顶部新增 `from vllm.config import ParallelConfig` 导入, 并将 `vllm_config.parallel_config` 替换为 `ParallelConfig()`, 让 fixture 携带完整的生产配置默认值, mock 只保留模型对象部分。
3. 设计取舍: 作者保留了 `draft_parallel_config` 仍为 `SimpleNamespace(tensor_parallel_size=1)`, 因为该字段未被 loader 读取扩展, 修复只针对实际契约缺口, 不过度扩大 mock 替换范围。
4. 验证配套: 本地运行该单测通过; pre-commit (Ruff、mypy、SPDX、配置默认值校验) 通过; Buildkite CI #89131 的 MI300 Model Executor lane 由 divakar-amd 确认修复。无配置、schema 或部署配套改动。

关键文件：

- tests/model_executor/test_qwen3_omni.py（模块 模型测试；类别 test；类型 test-coverage；符号 test_dspark_shares_target_embedding_with_smaller_draft_vocabulary）：唯一变更文件；修复 Qwen3 Omni DSpark embedding-sharing 单测中 parallel_config fixture 不完整导致的 CI 确定性失败，并把测试契约与生产 ParallelConfig 对齐。

关键符号：test_dspark_shares_target_embedding_with_smaller_draft_vocabulary

关键源码片段

tests/model_executor/test_qwen3_omni.py

唯一变更文件；修复 Qwen3 Omni DSpark embedding-sharing 单测中 parallel_config fixture 不完整导致的 CI 确定性失败，并把测试契约与生产 ParallelConfig 对齐。

```
@pytest.mark.skip_global_cleanup
def test_dspark_shares_target_embedding_with_smaller_draft_vocabulary():
    from vllm.v1.worker.gpu.spec_decode.dspark import utils as dspark_utils

    # 构造 target 与 draft 的 embedding 桩，draft 词表比 target 小 1，
    # 用于验证 DSpark 会复用 target embedding，而不是为 draft 单独建表
    target_embedding = nn.Embedding(100, 8)
    draft_embedding = nn.Embedding(99, 8)
    target_model = SimpleNamespace(model=SimpleNamespace(embed_tokens=target_embedding)
    )
    draft_model = SimpleNamespace(
        model=SimpleNamespace(embed_tokens=draft_embedding),
        has_own_embed_tokens=False,
    )
    draft_model_config = SimpleNamespace(
        hf_config=SimpleNamespace(model_type="qwen3"),
        get_vocab_size=Mock(return_value=99),
    )
    vllm_config = SimpleNamespace(
        speculative_config=SimpleNamespace(
            draft_model_config=draft_model_config,
            draft_parallel_config=SimpleNamespace(tensor_parallel_size=1),
            attention_backend=None,
            kv_cache_dtype=None,
        ),
        # 本次修复的关键一行：此前这里是 SimpleNamespace(tensor_parallel_size=1),
        # 而 #56387 之后 DSpark 的 _get_dspark_parallel_config 会继续读取
        # enable_eplb、eplb_config、enable_elastic_ep 等字段，缺字段导致
        # H200 / MI300 Model Executor CI 确定性失败；
        # 改用真实 ParallelConfig() 后，fixture 始终与生产配置契约保持一致
        parallel_config=ParallelConfig(),
        attention_config=SimpleNamespace(backend=None),
        cache_config=SimpleNamespace(),
    )
```

```

load_config=SimpleNamespace(),
model_config=SimpleNamespace(get_vocab_size=Mock(return_value=100)),
)

def fake_replace(config, **changes):
    # 模拟 dspark_utils.replace: 保留原字段并覆盖传入的变化
    values = vars(config).copy()
    values.update(changes)
    return SimpleNamespace(**values)

# 后续通过 patch 隔离 get_model、get_pp_group 等外部依赖,
# 再断言 draft 模型复用 target embedding, 此处从略

```

评论区精华

- divakar-amd 在 PR 评论区确认修复效果: "Thanks. This resolves (MI300) Model Executor (build #89113) ", 说明同一 fixture 缺陷同时影响了 H200 与 MI300 两条 lane, 该修复对两者均生效。
- khluu 触发 /ci run, github-actions 响应启动 Buildkite CI #89131, 最终验证结果与本地 pytest 一致。
- claude[bot] 提示这是 fork 提交的 PR, 自动审查被禁用; 维护者 njhill 直接批准合并, 无额外 review 评论。
- CI 触发与构建验证 (testing): 修复经 CI 验证有效, 覆盖 H200 与 MI300 两条 Model Executor lane, njhill 批准合并。
- fork PR 自动审查禁用 (other): 未额外运行 AI 审查, 由维护者人工确认后合并。

风险与影响

- 风险:
 - 测试行为开始依赖真实 ParallelConfig() 的构造逻辑: 若未来 ParallelConfig 在构造时引入环境敏感默认值 (如根据 CUDA_VISIBLE_DEVICES 推断 GPU 数) 或新增必填参数, 该测试可能再次出现环境相关不稳定; 当前默认值语义稳定, 风险低。
 - 该 fixture 被 DSpark embedding-sharing 测试复用, 本次只调整 parallel_config 字段, 不影响其它 mock 字段, 回归面窄。
 - 隐性契约风险仍在: 测试与 DSpark loader 读取字段之间仍是隐式耦合, 未来新增并行字段可能以 " 默认值与预期不符 " 的形式暴露, 但比 " 缺字段崩溃 " 更易诊断。
- 影响:
 - CI 稳定性: H200 与 MI300 Model Executor lane 不再确定性失败, 解除 #56387 引入的回归阻塞。
 - 开发者体验: Qwen3 Omni DSpark 测试 fixture 维护成本降低, 后续并行配置字段扩展不再需要逐字段修补。
 - 产品路径: 无运行时影响, 不触碰任何生产代码, 用户与部署无感知。
 - 风险标记: 测试夹具漂移, DSpark 配置契约, CI 回归

关联脉络

- PR #56387 [Bugfix][Spec Decode][MoE] Avoid uninitialized EPLB state in DeepSeek V4.1 DSpark drafter: 本 PR 的直接原因：它让 DSpark loader 从 target 的 parallel_config 读取 EPLB 字段，导致 Qwen3 Omni 测试 fixture 缺字段并在 CI 上确定性失败。
- PR #53458 [Bugfix][Spec Decode] Only create draft_id_to_target_id when draft vocab differs: PR body 明确提及它与最新失败构建相关但经排查无关，属于 Eagle3 词表映射的独立改动，本 PR 作者对此做了排除分析。
- PR #56903 [Perf][DSpark] Collapse DeepSeek-V4.1 draft states before SP all-gather: 同属 DSpark spec decode 模块 (vllm/v1/worker/gpu/spec_decode/dspark) 近期改动主线，说明该模块配置契约仍在演进，测试 fixture 需要持续对齐。