

PR #51410 完整报告

vllm-project/vllm

[CI] Refresh hybrid Model Runner V2 coverage

合并时间: 2026-08-08 13:10

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/51410>

执行摘要

- 一句话: 刷新 MRv2 混合模型 CI 覆盖并清理过时跳过
- 推荐动作: 值得 CI/ 测试维护者精读, 普通开发者可略读。亮点是系统性地清理 stale skip 并在多平台同步覆盖, 体现了对 MRv2 兼容性收口的推进; codex 提出的“定向用例应绑定模型源码依赖”是值得采纳的后续改进, 建议在后续 PR 中把 qwen3_5*.py 相关路径补进 source_file_dependencies。

功能与动机

PR body 明确列出四项动机: 移除 Qwen3.5 hybrid MTP 与 Jamba pipeline parallelism 下 stale 的 MRv2 skips, 增加 Qwen3.5 hybrid MTP 在 MRv2 下的定向 NVIDIA CI 覆盖, 修复 NVIDIA/Intel CI 中缺失的 synthetic rejection sampler 源码依赖, 以及用当前固定版本理由替换过时的 torchao nightly/0.12 注释。PR 同时说明使用 AI 辅助审计了 stale CI 覆盖, 并要求人工复核每一行改动后才能标记 ready。

实现拆解

1. 解除测试内的 stale skip: tests/v1/e2e/spec_decode/mtp/test_mtp.py 删除 import os, 并移除 test_mtp_correctness 函数中在 VLLM_USE_V2_MODEL_RUNNER=1 且模型名为 Qwen3.5 时的 pytest.skip(...) 分支。这样同一参数化用例 (ids 含 qwen3_5-hybrid) 在 MRv2 下恢复真实执行, 成为混合模型 V2 覆盖的载体。
2. 更新 NVIDIA 主 CI 配置: .buildkite/test_areas/model_runner_v2.yaml 中, Pipeline Parallelism job 的命令由 -k "not ray and not Jamba" 改为 -k "not ray", 恢复 Jamba 的 PP 覆盖; Spec Decode job 的 source_file_dependencies 增加 tests/v1/spec_decode/test_synthetic_rejection_sampler_utils.py, 并在 commands 末尾新增仅跑 test_mtp_correctness[qwen3_5-hybrid] 的定向命令。
3. 多平台同步: .buildkite/test-amd.yaml 的 PP job 同步去掉 and not Jamba; .buildkite/intel_jobs/model_runner_v2_intel.yaml 的 Spec Decode job 补上 synthetic rejection sampler 依赖, 保证 Intel CI 在该文件变更时也能被触发。
4. 注释清理: .buildkite/test_areas/quantization.yaml 用一行“Pin torchao to the version compatible with the CI PyTorch and CUDA stack.”替代 7 行过时的 nightly/0.12 注释, 安装命令 torchao==0.17.0 本身不变。

5. 验证与提交：作者通过 `/ci run` 触发了两轮 Buildkite (#82865、#82967)；提交仅一个功能 commit 加一次 merge main，没有新增单元测试，全部为 CI 配置与测试内跳过逻辑调整。

关键文件：

- `tests/v1/e2e/spec_decode/mtp/test_mtp.py`（模块 投机解码；类别 test；类型 test-coverage；符号 `test_mtp_correctness`）：核心变更：删除 MRv2 下对 Qwen3.5 hybrid MTP 的跳过逻辑和不再使用的 `os` 导入，使该参数化用例在 V2 runner 中真正执行。
- `.buildkite/test_areas/model_runner_v2.yaml`（模块 流水线配置；类别 config；类型 configuration）：MRv2 主 CI 配置：恢复 Jamba PP 覆盖、补齐 synthetic rejection sampler 依赖，并新增 Qwen3.5 hybrid MTP 定向用例。
- `.buildkite/test_areas/quantization.yaml`（模块 流水线配置；类别 config；类型 configuration）：清理过时的 torchao nightly/0.12 注释，替换为当前固定版本说明，避免误导后续维护者。
- `.buildkite/intel_jobs/model_runner_v2_intel.yaml`（模块 流水线配置；类别 config；类型 configuration）：Intel CI 的 Spec Decode job 补齐缺失的 synthetic rejection sampler 源码依赖。
- `.buildkite/test-amd.yaml`（模块 流水线配置；类别 config；类型 configuration）：AMD CI 的 MRv2 PP job 同步去掉 Jamba 过滤，与 NVIDIA 主 CI 保持一致。

关键符号：`test_mtp_correctness`

关键源码片段

`tests/v1/e2e/spec_decode/mtp/test_mtp.py`

核心变更：删除 MRv2 下对 Qwen3.5 hybrid MTP 的跳过逻辑和不再使用的 `os` 导入，使该参数化用例在 V2 runner 中真正执行。

```
# tests/v1/e2e/spec_decode/mtp/test_mtp.py（整理后片段）
# 关键变化：删除了原来基于 os.environ 判断 VLLM_USE_V2_MODEL_RUNNER 的 skip 块，
# 以及不再需要的 import os。这样在 V2 runner 下 qwen3_5-hybrid 参数会真实执行。
```

```
from typing import Any
```

```
import pytest
```

```
from tests.utils import single_gpu_only
from vllm import LLM, SamplingParams
from vllm.distributed import cleanup_dist_env_and_memory
from vllm.platforms import current_platform
```

```
from ..utils import (
    _skip_if_insufficient_gpus_for_tp,
    evaluate_llm_for_gsm8k,
    get_test_prompts,
)
```

```

@pytest.mark.parametrize(
    ["model_setup", "mm_enabled", "expected_accuracy_threshold"],
    [
        (("mtp", "XiaomiMiMo/MiMo-7B-Base", 1), False, 0.5), # ref: 65%-70%
        pytest.param(
            ("mtp", "ZixiQi/DeepSeek-V3-4layers-MTP-FP8", 1),
            False,
            0.0,
            marks=pytest.mark.skipif(
                current_platform.is_device_capability_family(100),
                reason="DeepSeek MTP: TRTLLM MoE top_k check fails on Blackwell",
            ),
        ), # dummy model
        # ... 其余参数化行 (qwen3_5-hybrid、gemma4-e4b) 保持不变
    ],
    ids=["mimo", "deepseek", "qwen3_5-hybrid", "gemma4-e4b"],
)

@single_gpu_only
def test_mtp_correctness(
    monkeypatch: pytest.MonkeyPatch,
    sampling_config: SamplingParams,
    model_setup: tuple[str, str, int] | tuple[str, str, int, str],
    mm_enabled: bool,
    expected_accuracy_threshold: float,
):
    """对比原模型与投机模型在 MTP 解码下的输出一致性,
    要求至少 6/10 输出 tokens 完全一致, 且 GSM8K 准确率高于阈值。
    """

    test_prompts = get_test_prompts(mm_enabled)
    with monkeypatch.context() as m:
        m.setenv("VLLM_MLA_DISABLE", "1")

        if len(model_setup) == 4:
            method, model_name, tp_size, draft_model = model_setup
        else:
            method, model_name, tp_size = model_setup
            draft_model = None
        _skip_if_insufficient_gpus_for_tp(tp_size)

        # 这里曾经是 MRv2 下对 Qwen3.5 hybrid 的 pytest.skip, 现已移除,
        # 因此在 VLLM_USE_V2_MODEL_RUNNER=1 时该用例会实际拉起模型并跑 GSM8K。
        attn_backend = "TRITON_ATTN" if current_platform.is_rocm() else "auto"

        # 跳过多模态 profiling: Qwen3.5 不启用 image/video, gemma-4 不启用 image/audio。
        extra_kwargs: dict[str, Any] = {}
        if "Qwen3.5" in model_name:
            extra_kwargs["limit_mm_per_prompt"] = {"image": 0, "video": 0}
        elif "gemma-4" in model_name:
            extra_kwargs["limit_mm_per_prompt"] = {"image": 0, "audio": 0}

```

```

if draft_model is not None and "gemma-4" in draft_model:
    import transformers
    from packaging.version import Version

    if Version(transformers.__version__) < Version("5.8.0"):
        pytest.skip(
            "Gemma4 MTP assistant requires transformers>=5.8.0, "
            f"got {transformers.__version__}"
        )

ref_llm = LLM(
    model=model_name,
    max_model_len=2048,
    tensor_parallel_size=tp_size,
    trust_remote_code=True,
    attention_backend=attn_backend,
    **extra_kwargs,
)
ref_outputs = ref_llm.chat(test_prompts, sampling_config)
evaluate_llm_for_gsm8k(
    ref_llm, expected_accuracy_threshold=expected_accuracy_threshold
)

```

评论区精华

唯一实质性 review 评论来自 codex 机器人，级别为 P2：新加的 Qwen3.5 hybrid MTP 定向用例所在 step 的 `source_file_dependencies` 只包含通用 GPU runner 路径和 spec test 路径，不含 `vllm/model_executor/models/qwen3_5.py`、`qwen3_5_mtp.py` 等模型文件；而 Qwen3.5 是默认走 V1 runner 的 hybrid 架构，普通 MTP suite 无法提供 V2 等价覆盖，因此模型侧改动可能在不触发该定向用例的情况下合入。建议把 Qwen3.5 模型、config 及 hybrid 层路径加入依赖。维护者 njhill 已 approve 并表示感谢，但该 P2 建议在合入版本中未见落实。另有 claude[bot] 说明 fork PR 自动 review 被禁用，chatgpt-codex-connector 只是触发提示。

- 新增 Qwen3.5 hybrid MTP 定向用例缺少模型源码依赖 (testing): 建议将 Qwen3.5 模型、config 及 hybrid 层路径加入 `source_file_dependencies`；维护者 approve 时未体现该修复，合入版本中依赖仍只有通用路径。

风险与影响

- 风险：
 1. CI 门禁暴露隐藏失败：移除 Qwen3.5 hybrid MTP 和 Jamba 的跳过过滤后，若这两个场景在 MRv2 下仍存在未修复问题，对应 NVIDIA/AMD CI job 会直接失败，相当于把隐藏故障提前暴露到主干门禁。
 2. 定向用例漏触发：codex P2 指出的依赖缺口仍然存在，`model_runner_v2.yaml` 的 Spec Decode step 未包含 `qwen3_5*.py` 等模型文件路径，模型侧回归可能无法触发新增的 `[qwen3_5-hybrid]` 用例。

3. 多平台行为一致性: AMD CI 的 PP job 仍是 # TBD 且 optional: true, Jamba 覆盖恢复后可能引入平台差异导致的失败; Intel CI 仅补了依赖声明, 未改变实际命令, 风险较低。
4. 注释变更无功能风险: quantization.yaml 只是注释清理, 安装命令不变, 风险几乎为零。
 - 影响: 对用户无运行时影响。对系统而言, MRv2 混合模型 (Qwen3.5 hybrid MTP、Jamba PP) 的 CI 覆盖从“跳过”变为“真实执行”, 回归发现窗口前移; AMD/Intel/NVIDIA 三平台配置一致性提升, synthetic rejection sampler 相关改动也能正确触发对应 job。对团队而言, NVIDIA H200 Job 新增一个定向用例, CI 耗时略有增加, 但避免了隐藏跳过带来的“假绿”误导; 后续需要跟进 codex 指出的模型源码依赖缺口。
 - 风险标记: CI 门禁可能暴露隐藏失败, 定向用例依赖缺失导致漏触发, 多平台 CI 配置同步, 无运行时代码变更

关联脉络

- PR #51413 [MRv2 Feature] MR v2 weight offloading support: 同为 MRv2 功能与覆盖扩展, 本 PR 继续刷新 MRv2 的 CI 覆盖范围。
- PR #51440 [CI Test] Add specific unit test for mrv2 offloading: 同一 MRv2 CI 收口方向, 都是为 MRv2 特定场景增加定向测试覆盖。
- PR #51310 [Spec Decode] Register Qwen3.6 dSpark acceptance coverage: 同为 spec decode 场景的 CI 覆盖扩展, 且都涉及 Qwen 系列模型的 e2e 测试注册。