PR #40472 完整报告

vllm-project/vllm

[CI] Add MTP coverage: Qwen3.5 correctness + no-sync spec decode

合并时间：2026-05-01 03:24

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/40472

执行摘要

- 一句话：为 Qwen3.5 MTP 新增 spec-decode CI 测试覆盖
- 推荐动作：建议合并。该 PR 针对测试矩阵的空白进行了精准补充，并且设计决策（阈值选择、兼容性跳过、视觉塔限制）都基于实测数据，具有一定参考价值。

功能与动机

PR body 明确了动机：'Adds hybrid-model coverage to two spec-decode tests that today only cover dense and MLA targets'。混合模型架构与密集 /MLA 有显著不同，缺乏覆盖导致已有 bug（#38556）无法被早期检测。Qwen3.5 是官方自带 MTP 权重的模型，适合用于填补这一空白。

实现拆解

1. 在 test_spec_decode.py 的 test_mtp_correctness 参数化列表中添加 ("mtp", "Qwen/Qwen3.5-0.8B-Base", 1) 案例，设置 GSM8k 阈值 0.20；增加 Model Runner V2 不兼容跳过；通过 limit_mm_per_prompt 限制视觉塔初始化 OOM。
2. 在 test_async_spec_decode.py 的 SPEC_DECODE_CONFIGS 中添加 self-draft 的 MTP 案例（mtp-qwen3_5-hybrid），同样限制视觉塔预热。
3. 在 spec_decode.yaml 中添加可选 B200 阶段 'Spec Decode MTP hybrid (B200)'，依赖 qwen3_5.py 和 qwen3_5_mtp.py 模型文件，通过 -k qwen3_5-hybrid 筛选测试。

关键文件：

- tests/v1/e2e/spec_decode/test_spec_decode.py（模块 推测解码；类别 test；类型 test-coverage；符号 test_mtp_correctness）：核心测试文件：添加了 Qwen3.5 混合模型的 MTP 正确性测试案例，包括 GSM8k 阈值设置和兼容性跳过逻辑。
- tests/v1/e2e/spec_decode/test_async_spec_decode.py（模块 推测解码；类别 test；类型 test-coverage；符号 test_no_sync_with_spec_decode）：异步无同步测试：添加了 Qwen3.5 MTP 案例，测试在混合模型下不会出现隐式 GPU-CPU 同步。
- .buildkite/test_areas/spec_decode.yaml（模块 CI 配置；类别 config；类型 configuration）：CI 配置：新增专门的 B200 通道用于运行 Qwen3.5 MTP 测试，明确定义源文件依赖。

关键符号：test_mtp_correctness, test_no_sync_with_spec_decode

关键源码片段

tests/v1/e2e/spec_decode/test_spec_decode.py

核心测试文件：添加了 Qwen3.5 混合模型的 MTP 正确性测试案例，包括 GSM8k 阈值设置和兼容性跳过逻辑。

```
# 在 test_mtp_correctness 的参数化列表中新增 Qwen3.5 案例
@pytest.mark.parametrize(
    ["model_setup", "mm_enabled", "expected_accuracy_threshold"],
    [
        (("mtp", "XiaomiMiMo/MiMo-7B-Base", 1), False, 0.5),
        (("mtp", "ZixiQi/DeepSeek-V3-4layers-MTP-FP8", 1), False, 0.0),
        # 新增混合模型案例：Qwen3.5-0.8B-Base (Mamba GDN + Attention)
        (
            ("mtp", "Qwen/Qwen3.5-0.8B-Base", 1),
            False,
            0.20, # 参考 GSM8k 精度 ~0.348，留约 14pt 余量
        ),
    ],
    ids=["mimo", "deepseek", "qwen3_5-hybrid"],
)

def test_mtp_correctness(...):
    ...
    # 跳过 Model Runner V2 不兼容情况
    if "Qwen3.5" in model_name and os.environ.get("VLLM_USE_V2_MODEL_RUNNER"):
        pytest.skip("Model Runner V2 does not yet support hybrid models")
    # 限制视觉多模态输入以防止 ViT 预热 OOM
    extra_kwargs: dict[str, Any] = {}
    if "Qwen3.5" in model_name:
        extra_kwargs["limit_mm_per_prompt"] = {"image": 0, "video": 0}
    # 构建参考 LLM 和 spec LLM 时传递 extra_kwargs
    ref_llm = LLM(model=model_name, ..., **extra_kwargs)
    ...
    spec_llm = LLM(model=model_name, speculative_config={...}, ..., **extra_kwargs)
```

评论区精华

评审者 benchislett 要求 GSM8k 阈值不能为零，必须基于实测设定合理值。作者参考测量后将阈值从 0.0 调整为 0.20。gemini-code-assist[bot] 建议在新 CI 通道的 source_file_dependencies 中加入具体模型实现文件，作者已采纳并提交补丁。

- GSM8k 正确性阈值应基于实测设定 (correctness): 阈值从 0.0 调整为 0.20（使用 Base 模型后），留出约 14pt margin，与 MiMo 案例一致的做法。
- 新 CI 通道应包含模型源文件依赖 (design): 作者在后续提交中添加了这两个文件依赖。

风险与影响

- 风险：主要风险来自新测试可能因未来 Qwen3.5 模型变更或 B200 驱动变化而失败，但测试为可选通道，不影响主 CI。另外，limit_mm_per_prompt 的 hack 假设视觉塔不需要，若后续测试需要多模态则失效。Model Runner V2 的跳过需跟踪上游修复。
- 影响：此 PR 全部是测试和 CI 配置变更，不涉及生产代码。对用户无直接影响。对 CI 系统，新增了一个可选的 B200 通道，在拥有该硬件的环境中可自动检测回归。对团队，提高了混合模型 spec-decode 的测试覆盖率，降低了未来修改导致错误的风险。
- 风险标记：硬件依赖 B200, ViT 内存限制 hack, Model Runner V2 兼容性跳过

关联脉络

- PR #38556 stale num_accepted_tokens_cpu corrupting hybrid hidden state under async spec decode: PR body 说明此 PR 新增的正确性测试本可捕获该 bug。
- PR #39546 B200 CI lane support for MTP hybrid: PR body 说明 B200 通道依赖此 PR，合入前需等待。