PR #42677 完整报告

vllm-project/vllm

[CI] Add MTP + PD disagg test for Qwen3.5

合并时间：2026-05-19 17:44

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42677

执行摘要

- 一句话：添加 MTP + PD 分解测试覆盖 Qwen3.5
- 推荐动作：本 PR 设计清晰，将测试配置与执行分离，便于后续扩展。建议关注测试结果，确保 MTP 在多种硬件配置下的稳定性。可以考虑后续添加更多 MTP 模型基线。

功能与动机

作为 PR #41869 的后续，本 PR 旨在为 Qwen3.5 的 MTP 方法在 PD 分解场景下提供验收测试覆盖，确保 MTP 在 NixlConnector 环境下也能达到预期的接受长度基准。

实现拆解

1. 重构测试数据模型：在 test_spec_decode_acceptance.py 中，将原有的 Eagle3ModelConfig 泛化为 ModelConfig，新增 method 字段以区分不同推测解码方法。添加 Qwen3.5-0.8B-Base 的 MTP 配置条目，并将查找函数 _get_model_config() 改为同时匹配模型名称和方法。
2. 新增配置扫描脚本：创建 config_sweep_spec_decode_test.sh，遵循已有的 config_sweep_accuracy_test.sh 模式，定义 EAGLE3 和 MTP 两组环境变量配置，循环调用 spec_decode_acceptance_test.sh 执行测试。
3. 增强 shell 测试脚本：在 spec_decode_acceptance_test.sh 中新增对 ENABLE_HMA_FLAG、VLLM_SSM_CONV_STATE_LAYOUT 和 VLLM_SERVE_EXTRA_ARGS 环境变量的支持，确保 MTP 所需的 HMA 和 SSM 布局参数能正确传递给 vllm serve 命令。
4. 更新 CI 定义：在 .buildkite/test_areas/disaggregated.yaml 中将原来的 spec_decode_acceptance_test.sh 直接调用替换为 config_sweep_spec_decode_test.sh，从而使 CI 自动同时运行 EAGLE3 和 MTP 测试。

关键文件：

- tests/v1/kv_connector/nixl_integration/test_spec_decode_acceptance.py（模块 接纳测试；类别 test；类型 test-coverage；符号 Eagle3ModelConfig, ModelConfig, _get_model_config）：核心测试文件，重构了 ModelConfig 数据类以支持多种推测解码方法，添加了 MTP 配置条目，是测试扩展的主要改动。
- tests/v1/kv_connector/nixl_integration/config_sweep_spec_decode_test.sh（模块 配置扫描；类别 test；类型 test-coverage）：新增的配置扫描脚本，遵循已有模式，定义了

EAGLE3 和 MTP 两组环境变量配置，循环调用测试脚本，是支持多方法测试的驱动力。

- tests/v1/kv_connector/nixl_integration/spec_decode_acceptance_test.sh（模块 测试驱动；类别 test；类型 test-coverage）：shell 测试脚本，新增对 HMA、SSM 布局、额外参数的支持，是 MTP 测试正确执行的关键配套。
- .buildkite/test_areas/disaggregated.yaml（模块 CI 配置；类别 config；类型 configuration）：CI 配置文件，将原来的直接调用替换为 sweep 脚本，是测试自动化触发的最后一环。

关键符号：_get_model_config

关键源码片段

tests/v1/kv_connector/nixl_integration/test_spec_decode_acceptance.py

核心测试文件，重构了 ModelConfig 数据类以支持多种推测解码方法，添加了 MTP 配置条目，是测试扩展的主要改动。

```
@dataclass
class ModelConfig:
    model: str # 模型名称，如 "Qwen/Qwen3.5-0.8B-Base"
    method: str # 推测解码方法，"eagle3" 或 "mtp"
    expected_acceptance_length: float # 期望的接受长度基准值
    drafter: str = "" # drafter 模型路径（EAGLE3 需要）
    expected_acceptance_lengths_per_pos: list[float] = field(default_factory=list) # 各位置接受率
    expected_acceptance_rate: float | None = None # 整体接受率（可选）
    id: str = "" # 配置标识
    rtol: float | None = None # 相对容差

# 基准配置列表（MT-Bench, 80 prompts, 256 tokens, temp=0）
MODEL_CONFIGS = [
    ModelConfig(
        model="meta-llama/Llama-3.1-8B-Instruct",
        method="eagle3",
        drafter="RedHatAI/Llama-3.1-8B-Instruct-speculator.eagle3",
        expected_acceptance_length=2.60,
        expected_acceptance_lengths_per_pos=[0.7296, 0.5208, 0.3545],
        id="llama3-8b-eagle3",
    ),
    ModelConfig(
        model="Qwen/Qwen3.5-0.8B-Base",
        method="mtp",
        expected_acceptance_length=1.798,
        id="qwen35-0.8b-mtp",
    ),
]
```

tests/v1/kv_connector/nixl_integration/config_sweep_spec_decode_test.sh

新增的配置扫描脚本，遵循已有模式，定义了 EAGLE3 和 MTP 两组环境变量配置，循环调用测试脚本，是支持多方法测试的驱动力。

```
#!/usr/bin/env bash
set -euo pipefail

# Sweep wrapper for spec decode acceptance tests.
# Runs spec_decode_acceptance_test.sh once per configuration.

SCRIPT="v1/kv_connector/nixl_integration/spec_decode_acceptance_test.sh"

# EAGLE3: Llama-3.1-8B-Instruct with EAGLE3 speculator.
eagle3_config="SD_METHOD=eagle3 MODEL_NAME=meta-llama/Llama-3.1-8B-Instruct SD_
MODEL=RedHatAI/Llama-3.1-8B-Instruct-speculator.eagle3 NUM_SPEC_TOKENS=3"

# MTP: Qwen3.5-0.8B-Base with hybrid SSM flags.
mtp_config="SD_METHOD=mtp MODEL_NAME=Qwen/Qwen3.5-0.8B-Base SD_MODEL=Qwen/
Qwen3.5-0.8B-Base NUM_SPEC_TOKENS=1 BLOCK_SIZE=32 MAX_MODEL_LEN=4096 VLLM_
SSM_CONV_STATE_LAYOUT=DS ENABLE_HMA_FLAG=1 KV_BUFFER_DEVICES=cuda"

configs=(
  "$eagle3_config"
  "$mtp_config"
)

for cfg in "${configs[@]}"; do
  local_cfg_parts=()
  read -r -a local_cfg_parts <<< "$cfg"
  echo "-> Running with: ${cfg}"
  if ! env "${local_cfg_parts[@]}" bash "${SCRIPT}"; then # 使用环境变量覆盖运行测试
    echo "Test failed for config: ${cfg}"
    exit 1
  fi
done

echo "All spec decode acceptance tests passed!"
```

评论区精华

本 PR 无人工 review 评论，自动机器人评论未提出修改建议。维护者 NickLucche 直接批准（LGTM）。

- 暂无高价值评论线程

风险与影响

- 风险：主要风险是新的 MTP 测试仅覆盖 Qwen3.5-0.8B-Base 单一模型，若后续增加更多 MTP 模型，需要同步更新配置。此外，HMA 和 SSM 布局参数的传递依赖于环境变量，存在变量未设置时默认行为不一致的风险。建议监控 CI 中出现的新测试失败。
- 影响：直接影响了 CI 流水线中 PD 分解测试的覆盖范围，新增了 MTP 方法的测试，并增加了测试运行时长度（约增加一倍）。对系统功能和性能无直接影响，属于测试基础设施增强。

- 风险标记：测试覆盖有限，参数传递兼容性，CI 耗时增加

关联脉络

- PR #41869 PD disaggregation with NIXL Connector: GDN support for Qwen3.5: 本 PR 是该 PR 的后续，为其添加 MTP 场景的 CI 测试覆盖。