PR #41294 完整报告

vllm-project/vllm

[ROCm][CI] Fix and stabilize EAGLE3 acceptance tests

合并时间：2026-06-02 01:40

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/41294

执行摘要

- 一句话：修复并稳定 ROCm 上 EAGLE3 测试
- 推荐动作：该 PR 值得精读，尤其是断言语义的调整和 EP 条件判断的设计，展示了如何在测试中平衡严格性与实用性。建议未来审视是否有其他 speculative decoding 测试需要类似处理。

功能与动机

ROCm CI 中 EAGLE3 接受长度测试存在不稳定：注意力后端硬编码 TRITON_ATTN 不兼容部分模型，gpt-oss 模型的接受率基线在 ROCm 上不同导致误判，Qwen3-VL FP8 模型在 TP=4 时因 block-FP8 分片形状不支持而崩溃，且双尾容差检查会将接受率改进视为失败。

实现拆解

PR 仅修改 tests/v1/spec_decode/test_acceptance_length.py 一个文件，共 5 个改动点：

1. 添加 ROCm 专属每位置接受基线：在 Eagle3ModelConfig 数据类中新增 rocm_expected_acceptance_lengths_per_pos 字段，并为 gpt-oss-20b-eagle3 配置设定 ROCm 下的基线 [0.7040, 0.4820, 0.3350]。
2. ROCm 注意力后端改为自动选择：修改 get_available_attention_backends，当平台为 ROCm 时返回 ["auto"] 而非之前的 ["TRITON_ATTN"]，让 vLLM 在运行时自动选择兼容后端。
3. 调整测试函数的后端处理：当 attention_backend 为 "auto" 时设置 attention_config=None，否则照常构建配置并检查排除。
4. 修复 Qwen3-VL TP=4 崩溃：通过启用 Expert Parallelism（enable_expert_parallel=True 当 tp==4 且模型为 Qwen3-VL），绕过 block-FP8 分片限制。
5. 改进接受率检查逻辑：将每个位置的检查从绝对相对误差（|actual-exp|/exp <= rtol）改为单边下限（actual >= exp*(1-rtol)），避免接受率提高时误报失败。同时为 get_mt_bench_prompts 添加必要参数以便测试多模态模型。

关键文件：

- tests/v1/spec_decode/test_acceptance_length.py（模块 测试；类别 test；类型 test-coverage；符号 Eagle3ModelConfig, get_available_attention_backends, test_eagle3_acceptance_length, get_mt_bench_prompts）：唯一变更文件，包含所有

ROCm 专属调整、注意力后端选择修改、EP 条件添加、断言逻辑改进。

关键符号：Eagle3ModelConfig, get_available_attention_backends, test_eagle3_acceptance_length, get_mt_bench_prompts

关键源码片段

tests/v1/spec_decode/test_acceptance_length.py

唯一变更文件，包含所有 ROCm 专属调整、注意力后端选择修改、EP 条件添加、断言逻辑改进。

```
# tests/v1/spec_decode/test_acceptance_length.py

@dataclass
class Eagle3ModelConfig:
    verifier: str
    drafter: str
    expected_acceptance_length: float
    expected_acceptance_lengths_per_pos: list[float] = field(default_factory=list)
    id: str = ""
    excluded_backends: set[AttentionBackendEnum] = field(default_factory=set)
    marks: list = field(default_factory=list)
    rtol: float | None = None
    # ROCm 专属配置：覆盖每个位置的期望接受率基线
    rocm_expected_acceptance_lengths_per_pos: list[float] = field(default_factory=list)

# ...（模型配置列表）
EAGLE3_MODEL_CONFIGS = [
    # 其他配置 ...
    Eagle3ModelConfig(
        verifier="openai/gpt-oss-20b",
        drafter="RedHatAI/gpt-oss-20b-speculator.eagle3",
        expected_acceptance_length=2.56,
        expected_acceptance_lengths_per_pos=[0.7165, 0.5120, 0.3337],
        id="gpt-oss-20b-eagle3",
        excluded_backends={AttentionBackendEnum.FLASHINFER},
        rocm_expected_acceptance_lengths_per_pos=[0.7040, 0.4820, 0.3350], # ROCm
        下基线不同
    ),
    # ...
]

def get_available_attention_backends() -> list[str]:
    # ROCm 平台使用自动后端选择，而非硬编码 TRITON_ATTN
    if current_platform.is_rocm():
        return ["auto"]

    get_valid_backends = getattr(current_platform.__class__, "get_valid_backends", None)
    if get_valid_backends is None:
```

```
        return ["FLASH_ATTN"]
    # ... 其余逻辑

def test_eagle3_acceptance_length(...):
    # 对于 "auto" 后端，不设置 attention_config，由 vLLM 自动选择
    attention_config = None
    if attention_backend != "auto":
        backend_enum = AttentionBackendEnum[attention_backend]
        if backend_enum in model_config.excluded_backends:
            pytest.skip(...)
        attention_config = {"backend": attention_backend}

    # 启用 Expert Parallelism 修复 TP=4 崩溃
    enable_ep = (tp_size == 4 and "Qwen3-VL" in model_config.verifier)

    with VllmRunner(...) as vllm_runner:
        ...
        # 求每个位置接受率的均值
        # 将检查从双尾改为单边：实际值必须 >= 期望值 * (1 - rtol)
        # 这样接受率提高时不会测试失败
        for pos, (actual, exp) in enumerate(zip(mean_acceptance_per_pos, expected_per_pos)):
            assert actual >= exp * (1 - rtol), \
                f"Per-position regression at pos {pos}: {actual} < {exp*(1-rtol)}"
```

评论区精华

- yewentao256 担心删除 TP=1 覆盖：micah-wil 证实 TP=1 实际可运行，失败原因是严格的双尾检查，因此通过改为单边检查保留覆盖。
- micah-wil 说明 TP=4 需要 Expert Parallelism 的原因，引用 issue #25292：block-FP8 在 TP=4 时因分片形状问题崩溃，开启 EP 后绕过。
- Bortlesboat 评论断言语义变化：从双尾转为单边，符合 speculative decoding 中更高接受率更好的直觉，但建议添加注释说明不对称性。同时指出总接受长度检查仍为双尾，可捕获全面向贪婪采样漂移的回归。
- TP=1 测试覆盖是否被删除 (correctness): micah-wil 确认 TP=1 实际可运行，之前失败是双尾检查过于严格，通过改为单边检查保留覆盖。
- TP=4 下 Qwen3-VL 需要 Expert Parallelism (correctness): 在测试中增加条件 enable_expert_parallel，当 tp_size==4 且模型包含 Qwen3-VL 时启用。
- 断言语义从双尾改为单边 (design): 接受，认为总接受长度检查仍为双尾，可兜底全面回归。

风险与影响

- 风险：
  - ROCm 后端改为 auto 选择：如果 auto 选择的后端表现不一致，可能引入不确定的测试基线偏移。
  - 单边检查可能掩盖接受率下降至 0 的严重回归（但总体检查仍保留双尾，提供兜底）。

    - Expert Parallelism 条件触发仅针对 TP=4 且模型含 "Qwen3-VL"，若其他模型同样需要 EP 则未被覆盖。
    - 测试覆盖率：该测试仅针对 EAGLE3，不影响生产代码。
- 影响：
    - 用户影响：仅影响 ROCm CI 测试稳定性，用户无需关注。
    - 系统影响：无。
    - 团队影响：减少 ROCm CI 误报和崩溃，提升开发效率。影响范围仅限于测试文件。
    - 风险标记：ROCm 后端 auto 选择不确定性，单边检查可能漏报极端回归，EP 条件可能不完整

关联脉络

- PR #44078 [MRV2] Remove Eagle's dedicated CUDA graph pool: 涉及 EAGLE3 speculative decoding 的 CUDA 图池移除，同属 EAGLE3 功能线，可能影响接受率测试。