PR #47920 完整报告

vllm-project/vllm

[Tests][Spec Decode] Add gemma4 MTP acceptance rates test

合并时间：2026-07-28 07:15

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/47920

执行摘要

- 一句话：新增 Gemma4 MTP 验收率测试并修复 3 个 bug
- 推荐动作：建议所有使用 Gemma4 MTP 的开发者 review 此 PR，特别是采样位置修复部分。测试框架的泛化设计值得学习，可复用于未来新增的推测模型。

功能与动机

Gemma4 MTP 缺乏验收率测试，同时存在三个运行时错误：

1) #43957 引入的 embedding 维度相等性检查阻碍了 Gemma4 MTP 的启动； 2) seq_lens 在 _prepare_decode_inputs_kernel 中未考虑 num_rejected，导致接受率次优； 3) 概率拒绝采样时 Gumbel 采样位置未随 draft 步数更新。本 PR 旨在填补测试空白并修复这些正确性问题。

实现拆解

1. 泛化验收率测试框架：将原有的 dflash_config fixture 和 test_dflash_acceptance_rates 合并为参数化的 test_acceptance_rates，新增 spec_config、expected_acceptance_lengths、chat_template_kwargs 三个参数，支持 DFlash 和 Gemma4 MTP 两种规格，并保留 MRV1/MRV2 双后端测试。
2. 修复 Gemma4 MTP embedding 共享：在 vllm/v1/spec_decode/gemma4.py 中覆盖 _maybe_share_embeddings 方法，绕过基类的 embedding 维度相等性检查，直接使用目标模型的 embed_tokens 替换 draft 模型的 placeholder embedding。
3. 修复 draft 解码 seq_lens 更新：在 vllm/v1/worker/gpu/spec_decode/autoregressive/speculator.py 的 _prepare_decode_inputs_kernel 中，将 seq_len = target_seq_len - num_rejected 计算移出 ADVANCE_DRAFT_POSITIONS 条件块，确保无论是否推进位置，seq_lens 都正确反映被拒绝的 token 数。
4. 修复 Gumbel 采样位置：在 _generate_draft 中，当 advance_draft_positions=False（Gemma4 MTP）时，计算 sample_positions = positions + current_draft_step 传递给采样函数，使得概率拒绝采样时 Gumbel 噪声在正确的绝对位置上生成。
5. 配套 CI 配置：在 .buildkite/test_areas/spec_decode.yaml 中新增 nightly 测试步骤 Spec Decode Acceptance Rates Nightly，在 H200 上运行 test_acceptance_rates。

关键文件：

- tests/v1/e2e/spec_decode/test_spec_decode.py（模块 规范解码测试；类别 test；类型 test-coverage；符号 dflash_config, test_dflash_acceptance_rates, test_acceptance_rates）：核心测试文件，重构原有 DFlash 测试为参数化框架，新增 Gemma4 MTP 验收率测试，覆盖 MRV1/MRV2 两后端。
- vllm/v1/spec_decode/gemma4.py（模块 Gemma4 MTP；类别 source；类型 dependency-wiring；符号 _maybe_share_embeddings）：修复 Gemma4 MTP 启动时的 embedding 共享问题，添加 _maybe_share_embeddings 覆盖方法。
- vllm/v1/worker/gpu/spec_decode/autoregressive/speculator.py（模块 推测核心；类别 source；类型 core-logic）：修复两个核心逻辑错误：seq_lens 更新未考虑 num_rejected、Gumbel 采样位置未偏移。
- .buildkite/test_areas/spec_decode.yaml（模块 CI 配置；类别 config；类型 configuration）：新增 nightly CI 步骤，确保验收率测试在 H200 上定期运行。

关键符号：test_acceptance_rates, _maybe_share_embeddings, _generate_draft, _prepare_decode_inputs_kernel

关键源码片段

vllm/v1/worker/gpu/spec_decode/autoregressive/speculator.py

修复两个核心逻辑错误：seq_lens 更新未考虑 num_rejected、Gumbel 采样位置未偏移。

```
# In _generate_draft method of Speculator class:
# 之前：draft_tokens = self.sample_draft(..., positions, ...)
# 之后：
sample_positions = positions
if not self.advance_draft_positions:
    # 当 draft 位置不推进（如 Gemma4 MTP 的 constant_draft_positions），
    # 采样位置需要累加 current_draft_step 以生成正确的 Gumbel 噪声。
    sample_positions = positions + self.current_draft_step

draft_tokens = self.sample_draft(
    last_hidden_states,
    sample_positions,
    idx_mapping,
    self.temperature,
    self.seeds,
    self.current_draft_step,
    self.draft_logits,
)
```

评论区精华

作者在 speculator.py 的 review 评论中说明了一个关键修复点：seq_len 应该无论是否推进 draft 位置都考虑 num_rejected，否则会导致 Gemma4 MTP 路径下 seq_lens 错误，进而影响接受率。该问题在 PR 中已修复。

- seq_len 应无条件考虑 num_rejected (correctness): 已修复：将 seq_lens 计算移出 ADVANCE_DRAFT_POSITIONS 条件块。

风险与影响

- 风险：
  - 回归风险：参数化测试覆盖了原有 DFlash 场景，基线值不变，但测试运行时间较长可能影响 CI 稳定性。
  - 性能风险：Gumbel 采样位置修复对吞吐的影响在 benchmark 中 < 2%，接受率持平或微升，无负面性能风险。
  - 兼容性风险：仅影响 Gemma4 MTP 推理路径，对其它模型或推测解码方法无影响。
  - 安全性风险：无。
- 影响：
  - 用户：Gemma4 MTP 用户将获得正确的接受率，概率采样不会因采样位置错误而退化。
  - 系统：新增 nightly 测试步骤，约 30 分钟，仅 CI 影响，不改变运行时。
  - 开发团队：为 spec decode 提供了可靠的回归保障，尤其是 Gemma4 路径。
  - 风险标记：核心路径变更，Gemma4 特定修复，采样逻辑修正，seq_lens 修复

关联脉络

- PR #43957 [Core] Add embedding-dim equality guard to _maybe_share_embeddings: 本 PR 修复了 #43957 引入的 embedding 维度检查导致的 Gemma4 MTP 启动问题。