PR #37639 完整报告

vllm-project/vllm

[Model Runner V2] Fix draft logits not populated during cudagraph replay

合并时间：2026-03-20 15:43

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/37639

执行摘要

- 一句话：修复 Eagle 投机解码在 CUDA 图重放时草稿 logits 未写入的问题，确保概率拒绝采样的正确性。
- 推荐动作：对于使用 Eagle 投机解码和 CUDA 图的工程师，建议精读此 PR，特别关注状态从 RequestState 移至 Speculator 的设计决策，以及 review 中关于数值精度的讨论。同时，可参考相关 PR 如 38045 以了解拒绝采样功能的更多上下文。

功能与动机

PR body 中指出：'When using probabilistic rejection sampling with Eagle speculative decoding and CUDA graphs enabled, the draft logits for speculative steps 1+ were not being written, causing incorrect rejection sampling behavior.' 根本原因是 draft logits 未传递给 EagleCudaGraphManager，因此未包含在 CUDA 图捕获中。修复后能提升系统准确性和性能，并为后续 PR 37588 的 CUDA 图全捕获功能铺路。

实现拆解

实现涉及三个文件的关键改动：1) model_runner.py：移除对 req_states.draft_logits 的依赖，在 sample 函数中改为使用 self.speculator.draft_logits，并更新 dummy_run 调用。2) speculator.py：在 EagleSpeculator 的 __init__方法中初始化 draft_logits 张量（基于配置决定缓存，使用 float32 类型），并更新 generate_draft 和 propose 方法以使用内部 draft_logits，移除相关参数。3) states.py：从 RequestState 的 __init__中彻底移除 draft_logits 初始化代码，简化状态管理。

关键文件：

- vllm/v1/worker/gpu/model_runner.py（模块 model_runner）：核心模型运行器文件，修改了 sample 和 dummy_run 函数，移除对 req_states.draft_logits 的依赖，改为使用 speculator.draft_logits，影响采样流程。
- vllm/v1/worker/gpu/spec_decode/eagle/speculator.py（模块 spec_decode/eagle）：Eagle 投机解码的核心实现，新增 draft_logits 存储并更新 init、generate_draft 和 propose 方法，是修复的关键所在。
- vllm/v1/worker/gpu/states.py（模块 states）：状态管理文件，从 RequestState 移除 draft_logits 初始化代码，简化结构并反映状态迁移。

关键符号：EagleSpeculator.init, EagleSpeculator.generate_draft, EagleSpeculator.propose, model_runner.sample

评论区精华

Review 中主要讨论了精度问题：gemini-code-assist[bot] 指出 draft logits 的 dtype 从 float32 改为 self.dtype（可能为 16 位浮点数）可能导致精度损失，影响概率拒绝采样的正确性。WoosukKwon 回复要求使用 float32，认为数值上重要且无足够证据降级为 bfloat16。此争议点已解决，PR 被批准，但凸显了 CUDA 图集成中对数值精度的敏感考量。

- 草稿 logits 数据类型精度问题 (correctness): WoosukKwon 要求使用 float32，认为数值上重要且无足够证据降级为 bfloat16，问题已解决。

风险与影响

- 风险：主要风险是精度损失：若 draft logits 使用非 float32 类型（如 float16），可能因下采样导致拒绝采样计算不准确，但 review 中已强制使用 float32 缓解此风险。其他风险包括回归风险，因为改变了状态存储位置（从 RequestState 移至 EagleSpeculator），可能影响依赖代码，但改动范围较小且测试通过；CUDA 图集成风险，需确保 draft logits 在重放时正确写入，但修复方案直接针对根因。
- 影响：对用户：修复后，Eagle 投机解码的概率拒绝采样将正确工作，提升接受率（如基准测试中从 42.44% 增至 45.79%）和输出吞吐量（如从 2,362.12 tok/s 增至 2,392.50 tok/s），改善生成质量。对系统：确保 CUDA 图重放时草稿 logits 被正确捕获，增强投机解码的稳定性和性能。对团队：代码结构更一致（匹配 draft_tokens 模式），为未来 CUDA 图功能（如 PR 37588）提供基础，降低维护成本。
- 风险标记：精度损失风险，CUDA 图集成风险

关联脉络

- PR #38045 [Model Runner V2] Enable forcing a specific acceptance rate during rejection sampling: 同属 Model Runner V2 的拒绝采样功能，标签 speculative-decoding，涉及相似模块和测试上下文。