PR #45672 完整报告

vllm-project/vllm

[Perf] Bound DiffusionGemma sampler transient via request-tiled logits

合并时间：2026-07-07 11:42

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/45672

执行摘要

- 一句话：分块限制 DiffusionGemma 采样器临时 logits 内存
- 推荐动作：值得精读。其动态分块策略设计精巧，可推广至其他需要控制临时内存的模型。建议关注未来是否添加相关测试。

功能与动机

采样器临时 logits 内存随 max_num_seqs 线性增长，在权重或 KV 池之前即 OOM。通过请求分块将瞬态内存从 O(num_decode * canvas * vocab) 限制为 O(group * canvas * vocab)。

实现拆解

1. 清除输出提前：将 sampled.zero_() 和 num_sampled.zero_() 从 _compiled_sample_step 函数内移到 Sampler.__call__ 的循环之前，避免分块循环中前一 tile 的结果被清零覆盖。
2. 自动计算分块大小：使用 torch.cuda.mem_get_info 获取当前 GPU 空闲内存，预算 50% 空闲内存用于采样器临时缓存，根据每请求所需字节数（约 canvas_length * vocab_size * 10）动态计算 group 大小。
3. 分块循环调用：在 Sampler.__call__ 中将 decode 请求按 group 分块，分别调用 _compiled_sample_step 处理，每个 tile 处理完后立即计算 logprobs 并释放 fp32 scaled logits 缓存，从而实现临时内存的降低。
4. 移除环境变量：根据 review 讨论，去掉新增的 VLLM_DIFFUSIONGEMMA_SAMPLER_TILE_SIZE 环境变量，改用上述自动分块策略。
5. 新增导入：添加 from vllm.platforms import current_platform，为后续平台相关特性做准备。

关键文件：

- vllm/model_executor/models/diffusion_gemma.py（模块 模型执行；类别 source；类型 core-logic；符号 _compiled_sample_step, Sampler.call）：唯一变更文件，包含采样分块、清零移动、内存预算计算等核心实现。

关键符号：_compiled_sample_step, Sampler.call

关键源码片段

vllm/model_executor/models/diffusion_gemma.py

唯一变更文件，包含采样分块、清零移动、内存预算计算等核心实现。

```
# vllm/model_executor/models/diffusion_gemma.py

# 新增导入：用于获取 GPU 平台特性
from vllm.platforms import current_platform

# 在 Sampler.__call__ 方法中，清空 sampled 和 num_sampled 被移出
# _compiled_sample_step 并放在循环之前，避免 tiling 时互相覆盖。
sampled = self._sampled[:num_reqs]
num_sampled = self._num_sampled[:num_reqs]
sampled.zero_()
num_sampled.zero_()

# 随后，根据 GPU 空闲内存动态计算分块大小并循环处理：
# group = max(1, int(free_mem * 0.5 / (canvas_length * vocab_size * 10)))
# for start in range(0, num_decode, group):
# # 调用 _compiled_sample_step 处理一个分块
# ...
```

评论区精华

Reviewer Isotr0py 反对引入模型专属环境变量，认为 DiffusonGemma 属于 dLLM 极端情况，建议基于空闲内存自动估计分块大小。作者接受并实现，使用 mem_get_info 自动计算组大小，并在 H200/H100/A100 等多种 GPU 上验证峰值内存保持稳定。

- 自动确定分块大小替代环境变量 (design): 作者采纳建议，移除环境变量，改用 torch.cuda.mem_get_info 自动计算组大小。

风险与影响

- 风险：
  1. 分块开销：分块循环引入了额外的 Python 循环和多次 GPU 内核启动，可能增加延迟。但 PR body 显示在 H200 上吞吐与原有单次调用持平。
  2. 自动内存计算：基于 mem_get_info 的分块大小可能受到系统内存碎片影响，但作者在多种 GPU 上验证峰值保持在 34-40% 内存，未 OOM。
  3. 无测试覆盖：本次变更未包含对应测试文件，回归风险较高，需依赖现有 e2e 测试。 - 影响：仅影响 DiffusionGemma 模型的采样逻辑，用户可在高并发场景下运行更大 batch size 而不 OOM。对系统其他模型无影响。兼容性完全保持，输出与单次调用 bit-identical。 - 风险标记：核心路径变更，缺少测试覆盖，模型特定优化

关联脉络

- PR #46155 [WIP] Fix DiffusionGemma sampler OOM with materialized logits (different approach): 作者 masterFoad 提出重叠问题的不同实现方案，本 PR 后续可能整合其验证数据或通用 hook。