PR #45417 完整报告

vllm-project/vllm

[Bugfix] Unset HF's default max_new_tokens for DiffusionGemma

合并时间：2026-06-15 17:31

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/45417

执行摘要

- 一句话：取消 DiffusionGemma 的 max_new_tokens 硬限制
- 推荐动作：简单明确的 bugfix，值得合并。无精读必要，但值得注意其设计思路：选择用 None 覆盖键值以跳过默认约束，而非直接修改 HF 配置或硬编码大值，保持了灵活性。

功能与动机

HuggingFace 的 DiffusionGemma generation_config.json 设置了 max_new_tokens=256，目的是防止用户在不指定 max_tokens 的情况下 OOM。但 vLLM 已有自己的内存管理机制，此硬限制会阻止模型生成超过 256 token 的多个 canvas。PR body 明确指出 'For vLLM implementation, we don't need to limit the max length by default in order to limit OOMs.'

实现拆解

1. 确定切入点：在 vllm/model_executor/models/config.py 的 DiffusionGemmaConfig.verify_and_update_config 方法中，已有对 DiffusionConfig 和 max_num_seqs 的调整逻辑，新代码附加在其后。
2. 判断并覆盖：检查 model_config.override_generation_config 字典中是否已存在 max_new_tokens 键（防止用户已通过 --override-generation-config 设置时被覆盖），若不存在，则将该键设为 None。设置为 None 的语义是让下游 get_diff_sampling_param 函数跳过该键，从而不使用 generation_config 中的默认值。
3. 日志提示：打印一条 info 日志告知用户这一行为，并提示可使用 --override-generation-config 自行设置上限。
4. 无需测试配套：由于是单点配置修复，且已有集成测试覆盖模型端到端行为，未添加新的测试文件。

关键文件：

- vllm/model_executor/models/config.py（模块 模型配置；类别 source；类型 data-contract）：唯一修改的文件，在 DiffusionGemmaConfig.verify_and_update_config 方法末尾添加了覆盖 max_new_tokens 的逻辑。

关键符号：verify_and_update_config

关键源码片段

vllm/model_executor/models/config.py

唯一修改的文件，在 DiffusionGemmaConfig.verify_and_update_config 方法末尾添加了覆盖 max_new_tokens 的逻辑。

```
# vllm/model_executor/models/config.py (DiffusionGemmaConfig.verify_and_update_config 末尾 )
# 前面已有 DiffusionConfig、max_num_seqs 的自动调整逻辑 ...

    # Remove the model's generation_config.json cap on max_new_tokens
    # (256) so DiffusionGemma behaves like every other model: no
    # server-wide limit, each request controls its own output length
    # via max_tokens. Setting to None causes get_diff_sampling_param
    # to skip this key entirely.
    model_config = vllm_config.model_config
    # 只有当用户未通过 --override-generation-config 显式设置时
    # 才覆盖为 None，避免覆盖用户意图
    if "max_new_tokens" not in model_config.override_generation_config:
        model_config.override_generation_config["max_new_tokens"] = None
        logger.info(
            "DiffusionGemma: removing server-wide max_new_tokens cap "
            "from generation_config.json (use "
            "--override-generation-config to set a custom limit).",
        )
```

评论区精华

无 review 评论。仅有的审核者 LucasWilkinson 快速批准，称 'LGTM, thanks for the fast follow!'

- 暂无高价值评论线程

风险与影响

- 风险：
  - 此变更仅影响 DiffusionGemma 模型，且仅在 override_generation_config 未包含 max_new_tokens 时执行赋值。若用户已通过 --override-generation-config max_new_tokens=100 显式设置，则不会覆盖，因此无副作用。
  - 风险极低，是标准的单文件 data-contract 修改。
- 影响：
  - 用户：现在可以生成超过 256 token 的 DiffusionGemma 输出，无需手动设置 max_tokens。
  - 系统：无额外性能开销，仅初始化时执行一次检查。
  - 团队：维护成本极低，是面向 DiffusionGemma 模型的一个小修补。
  - 风险标记：低风险

关联脉络

- 暂无明显关联 PR