PR #45429 完整报告

vllm-project/vllm

[Model] Support top_k and top_p sampling for DiffusionGemma

合并时间：2026-07-26 16:39

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/45429

执行摘要

- 一句话：DiffusionGemma 支持 top_k/top_p 采样
- 推荐动作：对于需要控制 DiffusionGemma 生成多样性的用户有益，建议合并。该 PR 还展示了处理 -inf 值与 zeroing 操作时数值稳定性的良好实践。

功能与动机

使用户能够像自回归模型一样，通过 top_k/top_p 参数控制 DiffusionGemma 的生成多样性。此功能在采样步骤之前过滤 logits，限制画布探索，而 committed argmax (top-1) 保持不变。

实现拆解

步骤如下：

1. 导入 apply_top_k_top_p 函数：从 vllm.v1.sample.ops.topk_topp_sampler 导入，复用现有的 top_k/top_p 过滤逻辑。
2. 在 __call__ 方法中，计算有效画布长度之后，根据 sampling_states.get_top_k_top_p 获取每个请求的 top_k/top_p 参数，若存在则对 logits 应用过滤。此操作在画布填充之前执行，使得填充位置的 logits 仍为均匀分布，不受过滤影响。
3. 将画布填充操作从 * valid 乘法改为 masked_fill_，因为过滤后的 -inf 值乘以 0 会产生 NaN，masked_fill_ 可避免该问题。

关键文件：

- vllm/model_executor/models/diffusion_gemma.py（模块 扩散模型；类别 source；类型 core-logic；符号 apply_top_k_top_p, DiffusionGemmaForCausalLM.call）：唯一的变更文件，核心实现添加 top_k/top_p 采样支持并修复数值稳定性问题。

关键符号：apply_top_k_top_p, DiffusionGemmaForCausalLM.call

关键源码片段

vllm/model_executor/models/diffusion_gemma.py

唯一的变更文件，核心实现添加 top_k/top_p 采样支持并修复数值稳定性问题。

```
# Per-request top_k/top_p, mirroring the AR sampler. Masked tokens
# become -inf and survive the temperature scaling in the compiled
# step, so Gumbel sampling, probs, and entropy all see the filtered
```

```
# distribution. The committed argmax (always the top-1 token) is
# unaffected; only the canvas exploration is constrained. Applied
# before canvas padding so phantom positions stay uniform.
if num_decode > 0:
    top_k, top_p = self.sampling_states.get_top_k_top_p(
        decode_slots.repeat_interleave(valid_canvas_len), decode_slots_np
    )
    if top_k is not None or top_p is not None:
        logits = apply_top_k_top_p(logits.float(), top_k, top_p)

# Pad any truncated canvas back to CL so the uniform-CL sampler math
# holds. Phantom (padded) positions are zeroed -> uniform logits -> high
# entropy (no premature convergence) and argmax 0 (stable); they are
# never committed (num_sampled == real length). masked_fill (not
# multiply) so -inf entries from top_k/top_p filtering above don't
# turn phantom rows into NaN.
if num_decode > 0 and valid_canvas_len_np.min() < CL:
    ar = torch.arange(CL, device=device)
    starts = valid_canvas_len.cumsum(0) - valid_canvas_len # row offset per req
    valid = ar.unsqueeze(0) < valid_canvas_len.unsqueeze(1) # [num_decode, CL]
    src = (starts.unsqueeze(1) + ar.unsqueeze(0)).clamp_max(logits.shape[0] - 1)
    logits = logits[src.reshape(-1)].masked_fill_(~valid.reshape(-1, 1), 0)
```

评论区精华

无 review 评论，审核者 Isotr0py 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：变更量小（18 行新增），仅影响 DiffusionGemma 模型的采样路径。使用已存在的工具函数，回归风险低。数值稳定性通过 masked_fill_ 修复得到保障。性能影响可忽略，因为过滤操作在编译步骤之外运行。
- 影响：用户可为 DiffusionGemma 传递 top_k 和 top_p 采样参数，控制生成文本的多样性。默认行为不变（不传参时跳过过滤）。对系统其他部分无影响。
- 风险标记：数值稳定性处理，小型变更

关联脉络

- PR #45163 Related Issue: 关联 Issue，DiffusionGemma 采样参数支持