PR #47332 完整报告

vllm-project/vllm

[Bugfix][Gemma4] Fix FA4 mm_prefix mask: add sliding window and absolute q_idx

合并时间：2026-07-05 08:46

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/47332

执行摘要

- 一句话：修复 FA4 后端 Gemma4 长上下文乱码
- 推荐动作：建议合入该 PR，它修复了 Gemma4 在 FA4 后端上的严重输出错误。推荐注意力后端开发者详细阅读 _make_mm_prefix_mask_mod 的修改，学习如何正确地将滑动窗口与自定义 mask_mod 结合，以及绝对位置计算在分块预填充中的关键作用。对于使用 mm_prefix 特性的模型维护者，可参考其滑动窗口 clamp 的设计。

功能与动机

根据关联 Issue #47300 和 PR body，Gemma4 视觉模型在 SM90（FlashAttention-4）后端上，当上下文长度超过滑动窗口（1024 tokens）且包含至少一张图像时，输出完全乱码（如重复的 '768368...' 或 'la la la...'）。Triton 后端在同一输入下输出正确，表明是 FA4 后端特有的 bug。根因在于 _make_mm_prefix_mask_mod 缺失滑动窗口约束和使用局部 q_idx，导致非图像 Token 的注意力分布被破坏。本 PR 旨在修复这两个根因，恢复长上下文多模态推理的正确性。

实现拆解

变更仅涉及一个文件 vllm/v1/attention/backends/flash_attn.py，按以下步骤实现：

1. 修改函数签名：在 _make_mm_prefix_mask_mod 中新增 sliding_window_left: int | None 参数，用于传递滑动窗口左边界大小（Triton 约定：1 + window_size[0]）；全局注意力层 ((-1,-1)) 传递 None。
2. 计算绝对 Q 位置：在 mask_mod 内部通过 seqlen_info 计算出 q_abs = q_idx + (seqlen_k - seqlen_q)，代替原来的局部 q_idx，使得因果比较、滑动窗口限制和 mm_prefix 范围判断都使用一致的绝对位置，与 Triton 参考路径一致。
3. 添加滑动窗口约束：当 sliding_window_left 不为 None 时，在 causal 条件上追加 (q_abs - kv_idx) < sliding_window_left，实现 (causal AND sliding_window) OR mm_prefix 语义。使用两条 @cute.jit 分支（有窗口和无窗口）避免 CuTe DSL 的运行时分支。
4. 调整调用端：在 FlashAttentionBackend.forward 中，根据 sliding_window_size 计算 sw_val = 1 + sliding_window_size[0]（当 window_size[0] >= 0 时），否则 None。该值同时用于 mm_clamp_sw（控制 mm_prefix 双向块的 clamp）和 sliding_window_left。mm_prefix_clamp_sliding_window 层属性（来自 PR #47217）决定是否启用 clamp。

5. 清理冗余：移除旧有的直接赋值和注释，统一使用 sw_val，提高可读性。

测试配套：本次改动未新增自动化测试文件，但作者在 PR body 中提供了详细的 NIAH（needle-in-a-haystack）手动复现步骤，并在 H100（SM90）上验证了修复结果。cjackal 在 Issue 评论中也确认了 dense 和 moe 模型均通过测试。

关键文件：

- vllm/v1/attention/backends/flash_attn.py（模块 注意力层；类别 source；类型 core-logic；符号 _make_mm_prefix_mask_mod, mm_prefix_mask_mod）：唯一修改文件，实现所有修复逻辑，包括函数签名变更、绝对位置计算、滑动窗口约束和调用端调整。

关键符号：_make_mm_prefix_mask_mod, mm_prefix_mask_mod

评论区精华

审核由 Isotr0py 完成并批准，未提出修改意见。cjackal 在关联 Issue #47300 中手动验证了修复在 dense 和 moe 模型上均有效（'Thanks for a quick fix! I have verified that this PR branch passes the test script without trouble for both dense and moe model.'）。Mergify 自动化提醒分支存在冲突，但在合并前成功解决。整体没有出现设计分歧或未解决疑虑。

- 审核与批准 (other): PR 获得批准，可以合并。
- 手动验证修复有效性 (testing): 验证通过，修复有效。

风险与影响

- 风险：变更严格局限在 _make_mm_prefix_mask_mod 一个函数内，仅影响 is_mm_prefix_lm=True 且使用滑动窗口的模型（目前主要是 Gemma4），风险可控：
  - 回归风险：全局注意力层（window=(-1,-1)）传递 sw_left=None，走无窗口分支，行为与修复前基本一致（仅 q_abs 修正可能带来微小变化，理论上修复错误，不引入退化）。
  - 手动验证覆盖有限：虽然 NIAH 测试涵盖了单块预填充（32k tokens）和分块预填充（144k tokens），但未覆盖所有序列长度和图像组合组合。
  - 硬件依赖：仅 SM90+ GPU（FA4）受影响，其他硬件（如 Triton 后端）不受影响。
  - 测试配套：无新增自动化测试，回归风险需由现有 CI 覆盖（可能不足）。
  - 影响：对用户的影响：之前使用 Gemma4 模型（H100 等 SM90+ GPU）进行长上下文视觉推理时输出完全乱码，修复后恢复正常，是严重 Bug 的关键修复。对于其他模型（如 bagel、molmo2、paligemma）或非 FA4 后端，行为不变。对系统影响：无性能退化，两条 @cute.jit 分支避免了运行时分支，应保持零开销。对团队影响：为后续 mm_prefix 模型正确集成滑动窗口提供了参考实现。
  - 风险标记：仅手动验证，无新增自动化测试，依赖特定 GPU 架构 (SM90+)，影响范围限于 FA4 + mm_prefix + sliding_window

关联脉络

- PR #47217 PR #47217: 引入 mm_prefix_clamp_sliding_window 层属性：本 PR 依赖该 PR 引入的 mm_prefix_clamp_sliding_window 标志来启用对 mm_prefix 双向块的滑动窗口 clamp，PR body 中明确引用。