PR #39444 完整报告

vllm-project/vllm

[Bugfix] Fix V1 dummy run writing NaN to KV cache null block

合并时间：2026-04-10 16:09

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39444

执行摘要

本 PR 修复了 V1 中 _dummy_run 函数将 NaN 写入 KV 缓存 null block 的 bug，该 bug 影响所有使用数据并行（DP）和专家并行（EP）的部署，导致模型精度显著下降（如 DeepSeek R1 从 97% 降至 55%）。修复通过填充 slot mapping 为 -1 使内核跳过 KV 写入，与 V2 行为对齐，已通过现场测试验证 NaN 消除。

功能与动机

问题背景：在 DP+EP 部署中，空闲的 DP rank 会执行 _dummy_run 以保持同步。由于 #25954 重构后，_dummy_run 无条件调用 _get_slot_mappings，而 slot mapping 缓冲区初始化为 torch.zeros，导致 slot_idx=0 映射到 KV 缓存的 null block（block 0），使 concat_and_cache_mla 内核写入 NaN。

影响：所有使用 DP+EP 的 V1 部署均受影响，无论量化类型（FP8、NVFP4、BF16），表现为精度严重回归。PR body 中提到在 DeepSeek R1 NVFP4 部署中观察到精度从 97% 跌至 55%。

实现拆解

主要改动集中在 vllm/v1/worker/gpu_model_runner.py 的 _dummy_run 函数中：

1. 参数更正：将 _get_slot_mappings 调用中的 num_tokens_padded 参数从 num_tokens（未填充计数）改为 num_tokens_padded（填充后计数），确保内部填充逻辑正确执行。 python slot_mappings_by_group, slot_mappings = self._get_slot_mappings(num_tokens_padded=num_tokens_padded, # 原为num_tokens ...)
2. slot mapping 填充：在获取 slot mapping 后，遍历 slot_mappings_by_group 中的所有张量，使用 fill_(-1) 设置为 PAD_SLOT_ID。 python if slot_mappings_by_group is not None: for sm in slot_mappings_by_group.values(): sm.fill_(-1) 这使 concat_and_cache 内核跳过 KV 写入，避免污染 null block。

评论区精华

review 中主要讨论了修复的完整性：

- gemini-code-assist[bot]指出："Filling the slot mappings with -1 is a correct approach... However, this fix appears to be incomplete due to an existing issue... _get_slot_mappings is called with num_tokens_padded=num_tokens, where

num_tokens is the unpadded count.”

- elvircrn回应：“I added this change num_tokens_padded=num_tokens_padded now as it seemed like a typo. But the reason why this wasn't an issue was because... slot_mapping[num_tokens_unpadded:num_tokens_padded].fill_(-1) already handles the rest of the data.”
- tlrmchlsmth最终批准修复。

风险与影响

风险分析：

- 回归风险低：修复针对性强，且通过现场测试验证（slot mapping 从 [0] 变为 [-1]，NaN 消除）。
- 性能影响可忽略：fill_(-1) 操作开销小，dummy run 本身频率较低。
- 兼容性良好：与 V2 的 get_dummy_slot_mappings 行为一致，无 breaking change。

影响评估：

- 用户：修复后模型推理准确性恢复，避免因 NaN 传播导致的错误输出。
- 系统：消除 KV 缓存污染，提升 DP+EP 部署的稳定性。
- 团队：揭示了 slot mapping 初始化与 dummy run 交互的设计缺陷，提示需加强相关测试覆盖。

关联脉络

- 历史 PR #25954：该 PR 重构了 slot mapping 逻辑，使 _dummy_run 无条件调用 _get_slot_mappings，引入了本 bug。
- 近期 PR 38794：同属 v1 worker 模块优化，涉及 GPU 模型运行器性能改进，可对比学习核心路径变更模式。
- 近期 PR 39169：同为 bugfix，涉及预热逻辑与真实路径对齐，展示了类似“修复不完整”讨论模式。

演进趋势：本 PR 是 V1 核心 worker 模块的持续稳定性改进的一部分，反映了团队对 DP+EP 等复杂并行场景下边缘案例的深入排查。