

PR #30001 完整报告

sgl-project/sglang

[NPU] bugfix for dsv4 memory pool

合并时间: 2026-07-03 16:41

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30001>

执行摘要

- 一句话: 为 NPU 添加空实现的 `clear_unaccepted_c128_draft_states`
- 推荐动作: 值得快速合并以修复 NPU 兼容性。建议未来添加注释说明为何留空, 方便后续维护。

功能与动机

PR #28612 引入了 `clear_unaccepted_c128_draft_states` 方法, NPU 路径也需要具备同名接口以避免 `AttributeError`, 即使 NPU 上无需实际清理。

实现拆解

在文件 `python/sglang/srt/hardware_backend/npu/dsv4/dsv4_memory_pool.py` 的 `NPUDeepSeekV4MemoryPool` 类中新增 `clear_unaccepted_c128_draft_states` 方法, 位于 `_make_indexer_state_pool` 方法之后。方法签名与 CUDA 版本一致, 但方法体仅为 `pass`。

关键文件:

- `python/sglang/srt/hardware_backend/npu/dsv4/dsv4_memory_pool.py` (模块 NPU 后端; 类别 source; 类型 core-logic; 符号 `clear_unaccepted_c128_draft_states`): 新增 `clear_unaccepted_c128_draft_states` 方法以匹配 CUDA 接口, 修复 NPU 上缺少方法定义的 bug。

关键符号: `clear_unaccepted_c128_draft_states`

关键源码片段

`python/sglang/srt/hardware_backend/npu/dsv4/dsv4_memory_pool.py`

新增 `clear_unaccepted_c128_draft_states` 方法以匹配 CUDA 接口, 修复 NPU 上缺少方法定义的 bug。

```
def clear_unaccepted_c128_draft_states(
    self,
    req_pool_indices: torch.Tensor,
    seq_lens: torch.Tensor,
    accept_lens: torch.Tensor,
    num_draft_tokens: int,
) -> None:
```

```
# NPU 的 fused compressor 不需要实际清理草稿状态,  
# 此方法仅提供空实现以匹配 CUDA 接口,  
# 避免调用时抛出 AttributeError。  
pass
```

评论区精华

无 review 讨论，PR 由 sglang-npu-bot 提交并合并。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。新增的方法为空实现（pass），不会改变现有行为。但如果未来 NPU 需要实际清理逻辑，空实现可能被遗忘更新。
- 影响：仅影响 NPU 硬件后端，修复了调用 `clear_unaccepted_c128_draft_states` 时抛出的 `AttributeError`。不影响 CUDA 或其他后端。
- 风险标记：空实现可能被遗忘

关联脉络

- PR #28612 Introduce `clear_unaccepted_c128_draft_states`: 本 PR 是对 PR #28612 引入的接口的 NPU 适配，原 PR 添加了 CUDA 版本的方法。