# PR #5768 完整报告

verl-project/verl

[trainer] feat: support use_remove_padding=False for mindspeed backend

合并时间：2026-03-28 15:34

原文链接：http://prhub.com.cn/verl-project/verl/pull/5768

## 执行摘要

本 PR 为 mindspeed（NPU）后端添加了在 use_remove_padding=False 配置下的支持，通过引入 NPU 特定的注意力掩码处理函数并修复实现 bug，提升了在 Ascend 设备上的训练兼容性。变更集中在模型工具文件，风险较低，但需关注设备特定逻辑的集成。

## 功能与动机

此变更旨在修复 issue #5704（具体内容未提供），解决 mindspeed 后端在 use_remove_padding=False 时可能因注意力掩码格式不匹配导致的运行时问题。PR body 中明确引用该 issue，动机是增强 NPU 后端的训练功能。

## 实现拆解

实现仅修改文件 verl/models/mcore/util.py，关键改动点如下：

- 新增函数 _build_npu_attn_mask：构建适用于 torch_npu.npu_fusion_attention 的注意力掩码（B1SS 形状）。代码逻辑包括生成因果掩码并与原始掩码组合。
- 修改 preprocess_bshd_no_padding 函数：在 NPU 可用时（通过 is_npu_available 检查），调用 _build_npu_attn_mask 转换注意力掩码，确保符合 NPU 后端要求。
- 修改 postprocess_bshd_no_padding 函数：在 NPU 可用时，还原注意力掩码格式，保持后处理一致性。
- 其他调整：修正了断言语句的拼写错误（从 'without bshd' 改为 'with bshd'）。

## 评论区精华

review 讨论仅包含一条来自 gemini-code-assist[bot] 的评论，聚焦于代码正确性：

> 评论指出 _build_npu_attn_mask 函数存在两个 bug：第 505 行张量解包错误和第 508 行未定义变量，并提供了修正建议。 讨论迅速解决，无其他争议，结论是 bug 修复后代码被批准。

## 风险与影响

风险分析：

- 初始实现中的 bug 已修复，但 NPU 特定逻辑可能在其他后端引入意外行为，依赖 is_npu_available 条件检查来隔离。

- 缺乏显式的单元测试覆盖 NPU 路径，从材料中未提及测试变更，可能增加未来回归风险。
- 变更影响核心模型预处理流程，若条件检查失效可能导致训练错误。

影响评估：

- 对用户：使用 NPU 后端且配置 use_remove_padding=False 的训练任务将获得更好兼容性，潜在提升性能。
- 对系统：变更范围有限，仅影响单个文件，不破坏现有功能。
- 对团队：作为 NPU 支持演进的一部分，需持续关注设备特定优化和测试覆盖。

## 关联脉络

从近期历史 PR 分析看，此 PR 是仓库对 Ascend（NPU）设备支持持续投入的一部分：

- PR #5756 在 Ascend 950 设备上启用 MXFP8 rollout，与本 PR 同属硬件优化方向。
- PR #5734 添加 NPU 夜间 CI，增强测试和验证基础设施。
- PR #5740 补充 NPU 依赖，确保环境兼容性。 这些 PR 共同推动 NPU 后端功能成熟，显示团队在扩展多设备支持上的系统化努力。