PR #5900 完整报告

verl-project/verl

[veomni] feat: bump veomni to v0.1.8

合并时间：2026-04-15 17:13

原文链接：http://prhub.com.cn/verl-project/verl/pull/5900

执行摘要

- 一句话：升级 VeOmni 至 v0.1.8，修复并行参数并新增打包序列 Flash Attention 预处理。
- 推荐动作：建议工程师精读此 PR，重点关注 _prepare_veomni_flash_attention_kwargs 函数的实现细节和设备处理，以及配置自动重写机制的设计决策，这些对理解 VeOmni 集成和序列并行优化有参考价值。

功能与动机

根据 PR body 和 Issue 讨论，升级 VeOmni 至 v0.1.8 以修复并行状态初始化和支持 Flash Attention。具体地，deerlu 在 Issue 评论中确认已升级 CI 中的 VeOmni 版本，以使用新特性如参数修复和 Flash Attention 优化。

实现拆解

1. 依赖版本升级：更新多个 CI 工作流文件（如 .github/workflows/e2e_ppo_trainer_veomni_vllm.yml），将 VeOmni 安装从 @v0.1.4 或 @v0.1.5 改为 @v0.1.8，并添加 --ignore-requires-python --no-deps 标志和固定 transformers 版本为 4.57.3，确保测试环境一致性。
2. 配置类增强：在 verl/workers/config/engine.py 中，为 VeOmniEngineConfig 添加 _mutable_fields 以允许 attn_implementation 可变，并在 __post_init__ 中自动将 flash_attention_2/3/4 重写为对应的 VeOmni SP-aware 变体（如 veomni_flash_attention_2_with_sp），通过日志记录变更。
3. 核心逻辑实现：在 verl/workers/engine/veomni/transformer_impl.py 中：
   - 导入 prepare_fa_kwargs_from_position_ids 工具。
   - 修改 parallel_state.init_parallel_state 调用，将 ep_size 参数替换为 extra_parallel_sizes 元组。
   - 更新 build_parallelize_model 调用中的 basic_modules，使用集合操作去重合并模块列表。
   - 新增 _prepare_veomni_flash_attention_kwargs 函数，支持 2D 和 3D 打包位置 ID 格式，调用 VeOmni 工具预计算 cu_seq_lens 和最大长度，并返回 Flash Attention 所需参数字典。
4. 测试与文档配套：调整 tests/special_e2e/sft/test_sft_engine_all.sh 中的输出信息，更清晰地标识 veomni 后端测试；其他 CI 工作流同步更新依赖版本以覆盖 veomni 相关测试。

关键文件：

- verl/workers/engine/veomni/transformer_impl.py（模块 VeOmni 引擎；类别 source；类型 core-logic；符号 _prepare_veomni_flash_attention_kwargs）：核心实现文件，包含参数修复、新函数添加和 basic_modules 优化，直接影响 VeOmni 引擎的并行初始化和 Flash Attention 预处理。
- verl/workers/config/engine.py（模块 引擎配置；类别 source；类型 configuration）：配置类文件，修改 VeOmniEngineConfig 以自动重写 attn_implementation，提升序列并行兼容性，并添加日志记录。
- .github/workflows/e2e_ppo_trainer_veomni_vllm.yml（模块 CI 流水线；类别 infra；类型 infrastructure）：CI 工作流文件，升级 VeOmni 依赖版本至 v0.1.8，并调整安装参数，确保测试环境一致性和新功能验证。
- tests/special_e2e/sft/test_sft_engine_all.sh（模块 SFT 测试；类别 test；类型 test-coverage）：测试脚本，微调输出信息以更清晰地标识 veomni 后端测试，辅助验证升级后功能。

关键符号：_prepare_veomni_flash_attention_kwargs

关键源码片段

verl/workers/config/engine.py

配置类文件，修改 VeOmniEngineConfig 以自动重写 attn_implementation，提升序列并行兼容性，并添加日志记录。

```
def __post_init__(self):
    super().__post_init__()
    assert self.strategy in ["veomni"], f"strategy {self.strategy} not supported"

    # 自动重写flash_attention实现为VeOmni序列并行感知版本，提升兼容性
    replacements = {
        "flash_attention_2": "veomni_flash_attention_2_with_sp",
        "flash_attention_3": "veomni_flash_attention_3_with_sp",
        "flash_attention_4": "veomni_flash_attention_4_with_sp",
    }
    if self.attn_implementation in replacements:
        new_impl = replacements[self.attn_implementation]
        logger.info(f"Replacing attn_implementation from '{self.attn_implementation}' to '{new_
        impl}'")
        self.attn_implementation = new_impl # 修改配置值以使用VeOmni优化版本
```

评论区精华

review 中，gemini-code-assist[bot] 指出 _prepare_veomni_flash_attention_kwargs 函数返回的张量可能设备不匹配，建议将 cu_seq_lens 等移动到与输入 position_ids 相同的设备。此建议可能已被采纳以确保正确性，但 PR 讨论中未显示具体修改。最终 PR 由 wuxibin89 批准合并。

- 设备匹配问题 (correctness): 建议可能被采纳以增强函数健壮性，但 PR 讨论未显示具体修改；最终 PR 已合并，推测问题已解决。

风险与影响

- 风险：主要风险包括：1) 设备不匹配风险：如果 prepare_fa_kwargs_from_position_ids 返回 CPU 张量而模型在 GPU/NPU 运行，可能导致运行时错误；2) 依赖兼容性：升级 VeOmni 至 v0.1.8 可能引入不兼容变更，影响现有训练流程；3) 配置重写副作用：自动重写 attn_implementation 可能干扰用户显式配置，需确保日志清晰。风险集中在 verl/workers/engine/veomni/transformer_impl.py 的核心路径。
- 影响：影响范围：使用 VeOmni 引擎进行序列并行训练的用户将受益于改进的 Flash Attention 支持和参数修复，提升训练性能和稳定性。影响程度中等：直接修改了引擎配置和核心预处理逻辑，但未改变高层 API；CI 测试更新确保覆盖 veomni 后端，保障持续集成可靠性。
- 风险标记：设备匹配风险，依赖升级风险，配置重写副作用

关联脉络

- PR #5935 [ci] chore: Add veomni npu ci test: 同涉及 VeOmni 引擎的 CI 测试更新，关联 veomni 功能验证和依赖版本管理。