

PR #7377 完整报告

verl-project/verl

[ci, hardware] fix: stabilize ROCm PPO trainer CI

合并时间: 2026-08-19 09:45

原文链接: <http://prhub.com.cn/verl-project/verl/pull/7377>

执行摘要

- 一句话: ROCm PPO CI 升级镜像并禁用 vLLM custom all-reduce
- 推荐动作: 可以快速阅读本 PR 以了解 ROCm CI 的坑位: 核心要点是 vLLM custom all-reduce 在 ROCm 上不可用, 以及镜像内依赖应作为唯一事实来源。若后续在 ROCm 上继续开展 CI 优化, 建议关注 `disable_custom_all_reduce` 配置在 vLLM 侧的实现, 并考虑在更多 ROCm 工作流程中复用 `VLLM_ROCM_OVERRIDE` 模式。

功能与动机

PR 指出 ROCm 上的 PPO trainer CI 不稳定, 根因是 custom all-reduce 在 ROCm 上存在 bug, 且旧镜像需要运行时替换 Megatron 依赖才能满足版本要求, 这种替换既脆弱又可能引入镜像内依赖冲突。PR 明确说明“currently custom all reduce has some bugs on ROCM, disable it for verl CI on rocm”, 希望通过升级到包含所需依赖的 ROCm 7.14 镜像 + 禁用 custom all-reduce 来稳定 CI。PR body 还声明已经检索过相关 open PR, 避免重复。

实现拆解

本 PR 只改动一个 CI workflow 文件, 实现分三步:

1. 升级容器镜像: 将 `.github/workflows/e2e_ppo_trainer_megatron_vllm_rocm.yml` 中三个 job (Megatron vLLM PPO、Megatron-Bridge 3D 并行、cleanup) 的 `container.image` 从 `verlai/verl:rocm702-vllm020-te210-20260518` 统一改为 `verlai/verl:rocm7.14_torch2.12_release_0724`。新镜像基于 ROCm 7.14, 且已内置所需版本的 Megatron 依赖, 因此不再需要 runtime 安装。
2. 移除运行时依赖替换: 删除 Megatron-Bridge 相关 job 中手工执行 `pip uninstall -y mbridge`、`pip install git+...Megatron-Bridge...`、`pip install git+...Megatron-LM...`、`pip install nvidia-modelopt` 的四行步骤, 让镜像内依赖成为唯一来源, 避免与镜像预装包冲突。
3. 注入禁用 custom all-reduce 的 Hydra override: 在两个训练 job 的 `env` 中新增 `VLLM_ROCM_OVERRIDE: "+actor_rollout_ref.rollout.engine_kwargs.vllm.disable_custom_all_reduce=True"`, 并在每个调用 `tests/special_e2e_ppo_trainer/run_function_reward.sh` 或 `tests/special_e2e/run_ppo_trainer_megatron.sh` 的命令末尾追加 `"${VLLM_ROCM_OVERRIDE}"` (共 6 处)。这样所有 vLLM 推理调用都会统一关闭 custom all-reduce。

4. 配套动作：无源码、测试或文档改动；PR body 说明 **pre-commit** 全部通过， workflow 由提交者在本机 ROCm 环境验证。

关键文件：

- [.github/workflows/e2e_ppo_trainer_megatron_vllm_rocm.yml](#)（模块 CI 工作流；类别 infra；类型 infrastructure）：唯一变更文件，集中体现了 ROCm CI 稳定化的三项改动：镜像升级、移除 runtime 依赖替换、禁用 vLLM custom all-reduce。

关键符号：未识别

关键源码片段

[.github/workflows/e2e_ppo_trainer_megatron_vllm_rocm.yml](#)

唯一变更文件，集中体现了 ROCm CI 稳定化的三项改动：镜像升级、移除 runtime 依赖替换、禁用 vLLM custom all-reduce。

```
# 每个 vLLM 训练 job 的 env 中注入 Hydra override。
# VLLM_ROCM_OVERRIDE 用于关闭 vLLM custom all-reduce,
# 因为当前 custom all-reduce 在 ROCm 上有 bug
# （注释内提到的 ROCm 为硬件平台名）。
env:
  HOME: "/root"
  VLLM_ROCM_OVERRIDE: "+actor_rollout_ref.rollout.engine_kwargs.vllm.disable_custom_all_reduce=True"

# 在 run 脚本末尾追加 override 参数，让所有训练调用统一生效。
run: |
  ray stop --force
  VAL_BEFORE_TRAIN=True TEST_FREQ=1 SAVE_FREQ=1 SAVE_HF_MODEL=True \
  VERL_EXP_NAME="qwen2.5-0.5b-function-reward-minimal-fsdp-size8" \
  bash tests/special_e2e/ppo_trainer/run_function_reward.sh "${VLLM_ROCM_OVERRIDE}"

# 原先 runtime 手工覆盖 Megatron 依赖的步骤被删除，
# 新镜像已内置所需版本，避免与镜像内依赖冲突。
```

评论区精华

该 PR 没有 review 评论，也没有 issue 评论。PR body 中值得注意的说明是：本变更由人类提交者完成并验证，AI 辅助用于分析失败原因和准备提交元数据，提交者对所有改动行负责。reviewer wuxibin89 直接 APPROVED，未留多余意见。

- 暂无高价值评论线程

风险与影响

- 风险：风险集中在 CI 基础设施层，不影响训练 / 推理业务代码：
 1. 禁用 custom all-reduce 的性能影响：在 ROCm 上关闭 vLLM custom all-reduce 后，all-reduce 会回退到默认实现，理论上可能降低吞吐，但 CI 只跑小规模 E2E 验证，影响有

限，且换取的是稳定性。

2. 镜像升级带来的依赖漂移：新镜像 rocm7.14_torch2.12_release_0724 的依赖与旧镜像不同，移除 runtime 覆盖后如果镜像内 Megatron 有未知问题，可能导致新的失败；不过这正是 PR 要消除的脆弱点。
3. 环境变量注入形式依赖脚本支持：VLLM_ROCM_OVERRIDE 作为额外参数传给 shell 脚本，要求 run_function_reward.sh、run_ppo_trainer_megatron.sh 能正确透传该参数；如果脚本未来重构参数解析方式，这个 override 可能静默失效。
4. 覆盖面有限：只修了 Megatron 与 vLLM 的 ROCm E2E 流程，其他 ROCm 相关流水线（如 sglang）未覆盖，后续若 custom all-reduce 在其余路径再出现类似 bug，仍需要类似处理。
 - 影响：影响范围：仅限于 .github/workflows/e2e_ppo_trainer_megatron_vllm_rocm.yml 中涉及的 ROCm MI300 CI 作业。对用户无直接 API 或行为影响；对团队而言，ROCm 上的 Megatron + vLLM PPO E2E 回归测试会更稳定，减少因 custom all-reduce 崩溃导致的误报。改动小、风险局部，但能提升 CI 可靠性。
 - 风险标记：仅 CI 配置变更，禁用 custom all-reduce 影响推理性能，镜像升级带来依赖漂移，override 依赖脚本透传参数

关联脉络

- PR #7444 [ci] chore: Update ci image: 同为更新 CI 镜像，体现仓库在持续同步各硬件平台的 CI 运行环境。
- PR #7293 [ci] chore: Update ci image: 更新 Ascend CI 镜像并调整依赖，与本 PR 的镜像升级思路一致。
- PR #7404 [ci] chore: Add npu ci env: 通过 CI 环境变量控制运行时行为，与本 PR 用 VLLM_ROCM_OVERRIDE 控制 vLLM 配置的模式类似。