PR #1915 完整报告

radixark/miles

fix(lora): honor --attention-backend in the bridge LoRA path

合并时间：2026-07-29 07:30

原文链接：http://prhub.com.cn/radixark/miles/pull/1915

执行摘要

- 一句话：LoRA 桥接路径透传 attention-backend
- 推荐动作：值得快速浏览，无需精读。这是一次教科书式的参数透传遗漏修复：通过对比 model_provider.py 与 bridge 路径的差异定位根因，并同步修正了 CI 中不可能成功的参数请求。可关注的设计点是 " 一行修复 + 测试语义对齐 " 的最小 diff 实践，以及作者自我 review 移除注释保持提交简洁的做法。若团队有类似 " 批量复制参数到 provider" 的代码，建议考虑加参数透传完整性校验或对照测试。

功能与动机

PR body 明确指出：LoRA runs silently ignore --attention-backend, _setup_lora_model_via_bridge never copies it onto the provider, unlike model_provider.py:78。Megatron 会把该字段映射为 NVTE_*_ATTN 环境变量，字段丢失后 TE 会自行选择调用者已经明确排除的后端，这正是 test_lora_qwen2.5_0.5B.py 在 fused-attention backward 上 nightly 崩溃的原因。另外 gpt-oss 使用 --softmax-type learnable，TE 2.17 下 FlashAttention 被禁用且无 fused kernel，只剩 UnfusedDotProductAttention 可用，原 CI 请求 fused 属于无效请求，因此改为 auto。

实现拆解

1. 变更入口：miles/backends/megatron_utils/bridge_lora_helpers.py 的 _setup_lora_model_via_bridge，该函数集中把 args 上的并行度、重计算、激活显存等配置复制到 Megatron Bridge 返回的 provider 对象上，最后调用 provider.finalize()。
2. 核心变更：在 provider.distribute_saved_activations = args.distribute_saved_activations 之后、provider.variable_seq_lengths = True 之前插入一行 provider.attention_backend = args.attention_backend。这样 Megatron 才能把用户显式选择的注意力后端翻译成 NVTE_FLASH_ATTN / NVTE_FUSED_ATTN / NVTE_UNFUSED_ATTN 等开关；不透传时 TE 只能走启发式，本次修复让该参数在 LoRA 路径上真正生效。
3. 测试配套：tests/e2e/megatron/model_scripts/test_gpt_oss_20b_moe_lora_ci.py 的 misc_args 中把 --attention-backend fused 改为 --attention-backend auto。这是请求与 TE 实际能力对齐的必要修正，否则该 E2E 在 --softmax-type learnable 下永远不可能成功。
4. diff 演进：最初提交还带了三行解释性注释，作者在 review 中自我要求移除（remove the comments），最终提交只保留单行赋值，形成最小 diff。本 PR 无新增 CLI 参数、无

schema 或构建配套改动，因为 args.attention_backend 早已存在。

关键文件：

- miles/backends/megatron_utils/bridge_lora_helpers.py（模块 LoRA 桥接；类别 source；类型 core-logic；符号 _setup_lora_model_via_bridge）：核心修复文件：在 _setup_lora_model_via_bridge 的 provider 配置块中补上 attention_backend 透传，与 model_provider.py 对齐，修复 LoRA 路径静默忽略 --attention-backend 的问题。
- tests/e2e/megatron/model_scripts/test_gpt_oss_20b_moe_lora_ci.py（模块 E2E 测试；类别 test；类型 test-coverage；符号 execute）：测试配套：把 gpt-oss 20B MoE LoRA CI 的 --attention-backend 从 fused 改为 auto，与 TE 2.17 + learnable softmax 下无 fused kernel 的实际能力对齐。

关键符号：_setup_lora_model_via_bridge, execute

关键源码片段

miles/backends/megatron_utils/bridge_lora_helpers.py

核心修复文件：在 _setup_lora_model_via_bridge 的 provider 配置块中补上 attention_backend 透传，与 model_provider.py 对齐，修复 LoRA 路径静默忽略 --attention-backend 的问题。

```
# bridge_lora_helpers.py —— LoRA 桥接路径中把训练参数透传给 Megatron provider 的配置块。
# 此处集中复制并行度、重计算等配置；此前漏掉 attention_backend，导致 Megatron 无法生成
# NVTE_FLASH_ATTN / NVTE_FUSED_ATTN 等开关，TE 只能靠启发式自选后端。
provider.tensor_model_parallel_size = args.tensor_model_parallel_size
provider.pipeline_model_parallel_size = args.pipeline_model_parallel_size
provider.expert_model_parallel_size = args.expert_model_parallel_size
provider.sequence_parallel = args.sequence_parallel
provider.context_parallel_size = args.context_parallel_size
provider.gradient_accumulation_fusion = args.gradient_accumulation_fusion
provider.recompute_granularity = args.recompute_granularity
provider.recompute_method = args.recompute_method
provider.recompute_num_layers = args.recompute_num_layers
provider.distribute_saved_activations = args.distribute_saved_activations
# 本次修复新增：与 model_provider.py 对齐，让用户显式选择的注意力后端真正生效。
provider.attention_backend = args.attention_backend
provider.variable_seq_lengths = True
provider.moe_token_dispatcher_type = "alltoall"
provider.moe_router_load_balancing_type = "none"
# DSA 注意力有独立字段，按 args 上的 dsa_attention_backend 单独下发，互不影响。
if hasattr(provider, "dsa_attention_backend"):
    provider.dsa_attention_backend = getattr(args, "dsa_attention_backend", "megatron")
provider.finalize()
```

评论区精华

作者 yushengsu-thu 在 diff hunk 上自我 review 留言 remove the comments，随后在回复中说明 Removed in 80df93674 — the PR is now the single assignment line.，将最初的三行

解释性注释删除，最终提交只剩单行赋值。guapisolo 评论 let's wait for CI pass and then merge，作者回复 passed 后，guapisolo 给出 LGTM. 并 APPROVED，Zhichenzzz 也 APPROVED。Copilot 评审因请求者配额耗尽未能生效，实际审查由作者自查与两位维护者人工完成。

- 移除随行注释，保持最小 diff (style): 最终提交只保留单行赋值 provider.attention_backend = args.attention_backend，注释全部移除。
- 等待 CI 通过后合入 (other): CI 通过后合入，guapisolo 与 Zhichenzzz 均 APPROVED，PR 已合并。
- Copilot 评审未生效 (other): 无评审输出，审查由作者自查与两位维护者人工完成。

风险与影响

- 风险：行为变更：LoRA 训练现在会真实遵循 --attention-backend，原本依赖 TE 启发式选择后端的任务，kernel 选择可能改变，需要留意训练速度与显存表现。兼容性：provider.attention_backend 属性依赖当前 Megatron Bridge 版本，虽然与 model_provider.py 用法一致，但升级 Bridge 后应回归验证。测试盲区：gpt-oss E2E 只覆盖 auto 分支，fused 分支在 TE 2.17 + learnable softmax 下不可用，qwen2.5 等其他模型的 fused LoRA 路径仍依赖 nightly 覆盖兜底。风险整体较低，因为改动是单行赋值且与既有主路径对齐。
- 影响：影响所有走 bridge LoRA 路径的训练任务：显式指定注意力后端的用户将看到参数真正生效，不再被静默丢弃；对依赖 TE 启发式选择的存量任务，kernel 选择可能发生变化。对团队而言，修复了 main nightly 上一个已知崩溃来源，并让 gpt-oss 20B MoE LoRA CI 的请求与硬件 /TE 能力对齐，避免永久性无效请求。改动面极小，风险可控。
- 风险标记：TE 后端选择行为变更，E2E 仅覆盖 auto 分支，依赖 Bridge 字段存在

关联脉络

- PR #1794 feat(multi-lora): enable and validate MoE expert adapters: 改动了同一文件 miles/backends/megatron_utils/bridge_lora_helpers.py，同属 LoRA bridge 功能线；本 PR 补上了该路径参数透传的遗漏。
- PR #1928 [fix] DSA indexer on Blackwell: send the DSA indexer wk unquantized: 同为 Megatron 后端注意力路径修复，与本 PR 中 dsa_attention_backend 分支相邻，体现对注意力后端配置一致性的持续修补。