PR #33875 完整报告

sgl-project/sglang

[diffusion] Fix 4/8-step distilled MiniMax-H3 Turbo LoRA merge

合并时间：2026-08-07 12:38

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/33875

执行摘要

- 一句话：修复 MiniMax-H3 Turbo LoRA 2D 权重合并崩溃
- 推荐动作：值得精读 slice_lora_b_weights 的 2D 分片实现，它展示了如何按 output_sizes/output_partition_sizes 对 fused 权重做 TP 切分；在引入新的 fused LoRA 格式时可复用该模式。整体改动小、文档完整，可作为 LoRA 兼容性修复的参考样例。

功能与动机

MiniMax-H3 全质量推理需要 20+ 步去噪，社区 larryvrh/MiniMax-H3-Turbo-Lora 将模型蒸馏为 4-10 步即可产出可用 T2VA，大幅降低服务延迟。PR body 指出 Turbo keys 与原生 H3 命名一致，唯一阻塞是 fused 2D lora_B 权重（如 qkv_proj (21504, 64)、fc1 (28672, 64)）在合并时触发 IndexError，因此需要修复 MergedColumnParallelLinearWithLoRA 的切片逻辑。

实现拆解

1. 核心修复：修改 python/sglang/multimodal_gen/runtime/layers/lora/linear.py 中 MergedColumnParallelLinearWithLoRA.slice_lora_b_weights。先按 B.dim() 分支——3D 路径保留原 B[:, start:end, :] 逻辑；新增 2D 路径遍历 base_layer.output_sizes 与 output_partition_sizes，以 row_offset 累计各输出 section 的行偏移，按 tp_rank 切出每个 section 的本地分片，最后 torch.cat 拼接。
2. 回归测试：在 test_lora_format_adapter.py 的 _run_all_tests 中追加 larryvrh/MiniMax-H3-Turbo-Lora 用例，验证 minimax_h3_turbo_4step.safetensors 在格式适配前后均被识别为 STANDARD，防止未来格式检测回归。
3. 文档更新：docs/cookbook/diffusion/MiniMax/MiniMax-H3.mdx 新增 “Turbo LoRA for few-step generation” 章节，给出 /v1/set_lora 与 --num-inference-steps 4/8 用法、ckpt500 推荐及 ComfyUI 剪枝图 LoRA 不兼容警告，并将后续章节顺延；compatibility_matrix.mdx 加入已验证 LoRA 条目；quantization.mdx 修正指向 cookbook 的锚点。

关键文件：

- python/sglang/multimodal_gen/runtime/layers/lora/linear.py（模块 LoRA 层；类别 source；类型 core-logic；符号 MergedColumnParallelLinearWithLoRA.slice_lora_b_weights）：核心修复，新增 2D fused 权重的 TP 分片路径，解决 MiniMax-H3 Turbo LoRA

合并崩溃

- python/sglang/multimodal_gen/test/unit/test_lora_format_adapter.py（模块 格式适配；类别 test；类型 test-coverage；符号 _run_all_tests）：新增 MiniMax-H3 Turbo LoRA 格式回归用例，防止格式检测退化
- docs/cookbook/diffusion/MiniMax/MiniMax-H3.mdx（模块 文档；类别 docs；类型 documentation）：新增 Turbo LoRA 使用章节、示例与不兼容警告，并顺延后续章节编号
- docs/docs/sglang-diffusion/compatibility_matrix.mdx（模块 文档；类别 docs；类型 documentation）：将已验证的 MiniMax-H3 Turbo LoRA 加入兼容矩阵
- docs/docs/sglang-diffusion/quantization.mdx（模块 文档；类别 docs；类型 documentation）：修正指向 cookbook 的锚点编号，避免文档引用失效

关键符号：MergedColumnParallelLinearWithLoRA.slice_lora_b_weights, _run_all_tests

关键源码片段

python/sglang/multimodal_gen/runtime/layers/lora/linear.py

核心修复，新增 2D fused 权重的 TP 分片路径，解决 MiniMax-H3 Turbo LoRA 合并崩溃

```
# slice_lora_b_weights: 根据 LoRA 权重维度选择 TP 切分方式
def slice_lora_b_weights(self, B: torch.Tensor) -> torch.Tensor:
    tp_rank = get_tp_rank()

    if B.dim() == 3:
        # diffusers 风格 adapter 把 Q/K/V（或 gate/up）堆叠为 3D 张量，
        # 保持原有按 section 的第一维切片逻辑。
        shard_size = self.base_layer.output_partition_sizes[0]
        start_idx = tp_rank * shard_size
        end_idx = (tp_rank + 1) * shard_size
        return B[:, start_idx:end_idx, :]

    # 原生 fused checkpoint（如 MiniMax-H3 Turbo）每层仅一张拼接 2D 矩阵，
    # 需按输出 section 偏移逐个做 TP 分片后再拼接。
    shards: list[torch.Tensor] = []
    row_offset = 0
    for full_size, part_size in zip(
        self.base_layer.output_sizes,
        self.base_layer.output_partition_sizes,
    ):
        local_start = tp_rank * part_size
        local_end = (tp_rank + 1) * part_size
        shards.append(B[row_offset + local_start : row_offset + local_end, :])
        row_offset += full_size
    return torch.cat(shards, dim=0)
```

评论区精华

维护者 mickqian 在 issue 评论中提出 “could we update the cookbook as well, inserting a new section about the LoRA support?”，作者 niehen6174 随后回复 “Done” 并在后续 commit 中补充了 cookbook 章节和兼容矩阵条目。除此之外没有其他 review 评论，设计取舍主要由 PR body 中的根因分析支撑。

- cookbook 补充 LoRA 支持章节 (documentation): 已新增 'Turbo LoRA for few-step generation' 章节及兼容矩阵条目，文档问题关闭。

风险与影响

- 风险：
  - slice_lora_b_weights 的 2D 分支依赖 base_layer.output_sizes 与 output_partition_sizes 的顺序一一对应。若未来某个 fused LoRA 的 section 排布与 base_layer 不一致，切分会错位且不会有显式报错。
  - 新增的测试只覆盖 LoRA 格式识别（LoRAFormat.STANDARD），没有覆盖 TP>1 时 2D 分片的数值正确性。多卡环境下 merged lora_B 的行偏移计算若出错，可能导致静默错误。
  - 3D 路径保持原逻辑，diffusers 风格 LoRA 不受影响；但 2D 分支的 torch.cat 在分片数量较多时会引入额外拷贝，对加载性能影响很小。
  - 文档章节编号顺延和锚点更新若不彻底，可能造成其他页面引用失效（本 PR 已修正 quantization.mdx 中的一处锚点）。
  - 影响：对用户：MiniMax-H3 用户在 --model-variant fl2va 下可直接加载 larryvrh/MiniMax-H3-Turbo-Lora，用 4/8 步替代 50 步，显著降低 T2VA 推理延迟。对系统：扩展了 MergedColumnParallelLinearWithLoRA 的兼容面，但 3D 路径未改动，对已有 diffusion LoRA（Wan、HunyuanVideo 等）无回归。 对团队：补充了 cookbook 和兼容矩阵，降低了社区适配该 LoRA 的试错成本；该 2D 分片模式也便于后续复用。
- 风险标记：权重切分顺序敏感，TP>1 数值未验证，文档锚点变更，测试仅覆盖格式识别

关联脉络

- PR #33849 [diffusion] gate fast VAE paths by quality: 同属 diffusion 模块近期优化，围绕具体模型 / 路径做兼容与性能改进，与本 PR 一样补充了文档与测试，但未触碰 LoRA 代码
- PR #33818 [diffusion] Generalize the FLUX.2 VAE decoder fast path to AutoencoderKL (Z-Image / FLUX.1) behind quality=high: 同为 diffusion 模型特定快速路径优化，与本 PR 的模式相似，共同完善 diffusion 少步 / 快速推理能力