PR #21664 完整报告

sgl-project/sglang

[diffusion] Fix Flux.2

合并时间：2026-03-31 14:14

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21664

执行摘要

本 PR 修复了 Flux.2 扩散模型在 Tensor Parallelism (TP) > 1 时因权重加载错误导致输出空白图像的问题。通过为 to_out 层添加自定义 weight_loader，确保与非连续输入布局匹配，仅影响多 GPU 场景，提升了模型推理正确性。

功能与动机

Bug 引入于 commit 281fe10b5（来自 PR #20137），该提交将 Flux2ParallelSelfAttention 中的 to_out 层改为 RowParallelLinear(input_is_parallel=True)，同时将 to_qkv_mlp_proj 改为 MergedColumnParallelLinear(gather_output=False)。这导致输入为 [attn_shard | mlp_shard] 的非连续布局，而 RowParallelLinear 默认 weight_loader 假设连续布局，造成权重与特征不匹配。PR body 中明确指出："root cause ... leading to a mismatch between weights and input features。"

实现拆解

修改集中在 flux_2.py 文件的 Flux2ParallelSelfAttention 类中：

- __init__方法更新：添加条件检查 if self.tp_size > 1: 并调用 self._patch_to_out_weight_loader()，同时更新注释说明输入布局。
- 新增 _patch_to_out_weight_loader 方法：定义自定义权重加载器，逻辑如下： 该方法根据 TP rank 从完整权重中选择 attention 和 MLP 的对应列并拼接，适配非连续输入。

评论区精华

Review 中没有评论，表明 PR 可能通过其他渠道（如内部审核或直接合并）处理，无技术交锋或设计权衡讨论。

风险与影响

- 技术风险：自定义 weight_loader 逻辑依赖 inner_dim 和 mlp_dim 计算，如果维度分配错误或 TP 配置变化，可能导致权重加载不正确；仅处理 TP>1 场景，可能遗漏其他并行模式（如数据并行）。
- 影响范围：修复直接提升 Flux.2 模型在 TP>1 下的输出质量，避免空白图像；对单 GPU 或 nvfp4 用户无影响；团队需注意后续并行线性层修改可能引入类似布局不匹配问题。

关联脉络

- 关联 PR #20137：该 PR 引入了 bug，改变了并行线性层类型，是本修复的根源。
- 扩散模型演进：结合近期历史 PR（如 #21383 支持 NPU ring attention、#20757 优化并行解码），可见仓库正持续增强扩散模型的多平台和并行能力，本 PR 是这一趋势中的关键 bugfix，确保并行场景下的稳定性。