PR #22059 完整报告

sgl-project/sglang

[diffusion] fix FLUX[1,2]

合并时间：2026-04-05 16:03

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22059

执行摘要

本次 PR 修复了 SGLang 中 FLUX.1 和 FLUX.2 扩散模型的 bug，通过启用 guidance_embeds=True 加载检查点权重并移除 guidance_scale 的错误缩放，对齐 HuggingFace Diffusers 实现。影响范围限于多模态生成模块，提升了生成准确性和一致性，建议团队关注条件缩放逻辑的设计权衡。

功能与动机

PR body 明确指出：FLUX.1-dev 和 FLUX.2-dev 检查点包含训练好的 guidance_embedder 权重，但 SGLang 在 guidance_embeds=False 时丢弃这些权重。设置 guidance_embeds=True 是正确的修复，然而 _build_guidance() 函数错误地将 guidance_scale 乘以 1000，导致模型的正弦时间步嵌入层接收分布外的输入。Diffusers 直接传递原始 guidance_scale 值，因此需要移除缩放以匹配其规范。

实现拆解

- 编码器配置模块：新增 flux_2.py 定义 FLUX.2 的 Mistral 文本编码器配置，包括 build_flux2_text_messages 函数和 Flux2MistralTextConfig 类；在 flux.py 中移除冗余的 Flux2MistralTextArchConfig 和 format_text_input，改用 build_flux2_text_messages。
- 采样参数模块：更新 flux.py 中的 sampling params，修正默认值：FluxSamplingParams.guidance_scale=3.5，新增 Flux2SamplingParams.guidance_scale=4.0；在 registry.py 中注册配置时使用 Flux2SamplingParams。
- 运行时逻辑模块：在 denoising.py 中修改 _build_guidance() 函数，通过判断 pipeline_config 类型，仅对 FLUX.1 模型保持 *1000 缩放： python if isinstance(self.server_args.pipeline_config, FluxPipelineConfig) and not isinstance(self.server_args.pipeline_config, Flux2PipelineConfig): guidance_val = guidance_val * 1000.0 在 text_encoding.py 中调整 attention_mask 处理，区分 FLUX.1 和 FLUX.2 的编码流程。
- 加载器模块：在 component_loader.py 中将 Flux2PipelineConfig 的 tokenizer 加载从 AutoTokenizer 改为 AutoProcessor 以对齐 Diffusers。
- 测试模块：在 test_sampling_params.py 中添加测试验证 guidance_scale 默认值匹配模型需求。

评论区精华

review 中仅有 gemini-code-assist[bot] 的一条评论：

> "This pull request removes the 1000.0 scaling factor from the guidance tensor construction in the denoising stage to align with the HuggingFace Diffusers convention and prevent out-of-distribution embeddings." 无其他讨论，变更直接通过，表明团队对对齐 Diffusers 的共识。

风险与影响

- 风险：1) 回归风险：修改默认值和缩放逻辑可能影响 FLUX 模型生成质量，需准确性测试保障；2) 兼容性：条件缩放逻辑需精确区分 FLUX.1/FLUX.2，否则可能导致嵌入计算错误；3) 测试覆盖：新增配置和逻辑变更缺乏充分单元测试，可能隐藏 bug。
- 影响：范围限于扩散模型生成管线，程度中等。修正后提升生成准确性，确保与 Diffusers 一致性，对用户提供更可靠图像生成，对团队展示代码重构和对齐最佳实践。

关联脉络

从近期历史 PR 看，

- 22157 "[CI] No diffusers backend in lora case"：同属扩散模型修复，涉及 CI 测试和与 Diffusers 的一致性，共享多模态生成模块。
- 22146 "Isolate spec V1 path in decode post-processing"：类似结构重构模式，隔离不同版本路径，与本次 PR 中区分 FLUX.1/FLUX.2 逻辑有共通设计思路，反映仓库在演进中注重版本管理和一致性维护。