PR #25661 完整报告

sgl-project/sglang

[diffusion] model: support FLUX.2-klein-base

合并时间：2026-05-22 11:24

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25661

执行摘要

- 一句话：支持 FLUX.2-klein-base 未蒸馏模型，启用 CFG 和 negative prompts
- 推荐动作：该 PR 实现清晰，适合快速合并。建议关注其后的扩散模型 PR 以了解 FLUX 系列支持的演进。

功能与动机

此前 FLUX.2-klein-base 通过蒸馏版 Klein 间接支持，但蒸馏版不支持 negative prompts 和 CFG parallelism。本 PR 直接支持该模型，使其功能完备。

实现拆解

1. 新增 PipelineConfig：在 python/sglang/multimodal_gen/configs/pipeline_configs/flux.py 中添加 Flux2KleinBasePipelineConfig，继承自 Flux2KleinPipelineConfig，设置 should_use_guidance=True 并实现 prepare_neg_cond_kwargs 方法，为 CFG 路径构建 freqs_cis。
2. 新增 SamplingParams：在 python/sglang/multimodal_gen/configs/sample/flux.py 中添加 Flux2KleinBaseSamplingParams，设置默认 num_inference_steps=50、guidance_scale=4.0、negative_prompt=""。
3. 模型注册：在 python/sglang/multimodal_gen/registry.py 中导入新类，新增 register_configs 调用，并调整原始 Klein 的 detector 逻辑，增加 and "base" not in hf_id.lower() 以区分 base 和非 base 变体。
4. 放宽 negative_prompt 验证：在 python/sglang/multimodal_gen/runtime/pipelines_core/stages/text_encoding.py 中，将 CFG 的 negative_prompt 验证从 V.string_not_none(x) 改为 V.string_not_none(x) or isinstance(x, str)，允许空字符串（klein-base 的默认空提示）。
5. 测试与性能基线：在 python/sglang/multimodal_gen/test/server/gpu_cases.py 中添加了 1-GPU CI 测试用例，并在 perf_baselines.json 中记录实测性能数据。同时更新了文档中的兼容性矩阵。

关键文件：

- python/sglang/multimodal_gen/configs/pipeline_configs/flux.py（模块 扩散模型；类别 source；类型 core-logic；符号 Flux2KleinBasePipelineConfig, prepare_neg_cond_kwargs）：核心：新增 Flux2KleinBasePipelineConfig 及

prepare_neg_cond_kwargs 方法，控制 CFG 行为

- python/sglang/multimodal_gen/configs/sample/flux.py（模块 扩散模型；类别 source；类型 core-logic；符号 Flux2KleinBaseSamplingParams）：定义 Flux2KleinBaseSamplingParams，指定默认推理步数、guidance 和负提示
- python/sglang/multimodal_gen/registry.py（模块 注册中心；类别 source；类型 core-logic；符号 Flux2KleinBasePipelineConfig, Flux2KleinBaseSamplingParams）：注册新模型并调整 detector 逻辑，确保 base 和非 base 正确分流
- python/sglang/multimodal_gen/runtime/pipelines_core/stages/text_encoding.py（模块 文本编码；类别 source；类型 core-logic）：放宽 negative_prompt 验证，允许空字符串，确保 Klein-base 默认值通过
- python/sglang/multimodal_gen/test/server/perf_baselines.json（模块 性能基线；类别 test；类型 test-coverage）：新增 FLUX.2-klein-base 的性能基线数据
- python/sglang/multimodal_gen/test/server/gpu_cases.py（模块 GPU 测试；类别 test；类型 test-coverage）：添加 1-GPU CI 测试用例验证 Klein-base 推理
- docs_new/docs/sglang-diffusion/compatibility_matrix.mdx（模块 文档；类别 other；类型 documentation）：更新兼容性矩阵文档
- docs_new/docs/sglang-diffusion/dynamic_batching.mdx（模块 文档；类别 other；类型 documentation）：更新动态批处理文档中的模型列表
- .github/workflows/diffusion-ci-gt-gen.yml（模块 CI；类别 infra；类型 infrastructure）：CI 配置增加 Klein-base 测试步骤
- docs/diffusion/compatibility_matrix.md（模块 文档；类别 docs；类型 documentation）：旧版文档同步更新兼容性矩阵

关键符号：Flux2KleinBasePipelineConfig.prepare_neg_cond_kwargs

关键源码片段

python/sglang/multimodal_gen/configs/pipeline_configs/flux.py

核心：新增 Flux2KleinBasePipelineConfig 及 prepare_neg_cond_kwargs 方法，控制 CFG 行为

```
@dataclass
class Flux2KleinBasePipelineConfig(Flux2KleinPipelineConfig):
    # Undistilled Klein base model, with guidance embeddings
    should_use_guidance: bool = True

    def prepare_neg_cond_kwargs(self, batch, device, rotary_emb, dtype):
        # 获取负 prompt 的文本序列长度
        txt_seq_lens = self.require_text_seq_lens(
            batch,
            0,
            negative=True,
            expected_batch_size=batch.negative_prompt_embeds[0].shape[0],
        )
        # 为负 prompt 构建 rotary embedding 频率，用于 CFG 并行
```

```
        return {
            "freqs_cis": self.get_freqs_cis(
                batch.negative_prompt_embeds[0],
                batch.width,
                batch.height,
                device,
                rotary_emb,
                batch,
                txt_seq_lens,
            )
        }
```

python/sglang/multimodal_gen/configs/sample/flux.py

定义 Flux2KleinBaseSamplingParams，指定默认推理步数、guidance 和负提示

```
@dataclass
class Flux2KleinBaseSamplingParams(FluxSamplingParams):
    # Klein-base 是未蒸馏版本，需要 50 步和较大的 guidance
    num_inference_steps: int = 50
    guidance_scale: float = 4.0
    negative_prompt: str = "" # 允许空字符串，CFG 验证通过
```

python/sglang/multimodal_gen/registry.py

注册新模型并调整 detector 逻辑，确保 base 和非 base 正确分流

```
# 导入新配置
from sglang.multimodal_gen.configs.pipeline_configs.flux import (
    Flux2KleinBasePipelineConfig,
    Flux2KleinPipelineConfig,
    Flux2PipelineConfig,
)
from sglang.multimodal_gen.configs.sample.flux import (
    Flux2KleinBaseSamplingParams,
    Flux2KleinSamplingParams,
    Flux2SamplingParams,
    FluxSamplingParams,
)

# 注册蒸馏版 Klein，明确排除 base
register_configs(
    sampling_param_cls=Flux2KleinSamplingParams,
    pipeline_config_cls=Flux2KleinPipelineConfig,
    hf_model_paths=[
        "black-forest-labs/FLUX.2-klein-4B",
        "black-forest-labs/FLUX.2-klein-9B",
    ],
    model_detectors=[
        lambda hf_id: (
            "flux.2-klein" in hf_id.lower() or "flux2-klein" in hf_id.lower()
```

```
        )
        and "base" not in hf_id.lower() # 排除 base 变体
    ],
)

# 注册未蒸馏 Klein-base
register_configs(
    sampling_param_cls=Flux2KleinBaseSamplingParams,
    pipeline_config_cls=Flux2KleinBasePipelineConfig,
    hf_model_paths=[
        "black-forest-labs/FLUX.2-klein-base-4B",
        "black-forest-labs/FLUX.2-klein-base-9B",
    ],
    model_detectors=[
        lambda hf_id: (
            "flux.2-klein" in hf_id.lower() or "flux2-klein" in hf_id.lower()
        )
        and "base" in hf_id.lower() # 仅匹配 base
    ],
)
```

评论区精华

维护者 mickqian 要求作者提供结果与官方 diffusers 输出的对比，作者在 Issue 评论中附带了多组图片对比，显示输出质量一致。之后 mickqian 表示“fantastic job, cheers”并批准合并。无其他争议。

- 结果验证 (correctness): 作者提供了多组图片对比，显示输出质量一致，mickqian 表示满意。

风险与影响

- 风险：风险较低。主要变更在配置层和注册层，不涉及模型 forward 逻辑或 kernel 修改。潜在风险：模型 detector 逻辑修改可能影响非 base 版 Klein 的匹配（已通过精确排除 "base" 字符串控制）；empty string 验证放宽可能使其他模型传入空提示时通过验证，但这与 CFG 设计一致。
- 影响：对用户：新增对 FLUX.2-klein-base 的官方支持，可正常使用 CFG 和 negative prompts。对系统：无性能回归，新增测试用例和基线。对团队：约 150 行新增代码，维护成本低。
- 风险标记：新模型注册，测试覆盖完整，性能基线更新

关联脉络

- 暂无明显关联 PR