PR #21407 完整报告

sgl-project/sglang

[FIX] Flux2-Klein prompt tokenization length to 512 and add regression coverage

合并时间：2026-03-28 17:28

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21407

执行摘要 本次 PR 修复了 Flux2-Klein 扩散模型中提示词标记化长度错误的问题，将 max_length 从 77 改为 512，防止长提示被截断，并添加回归测试确保未来不会回归。

功能与动机 Issue #21372 报告了 Flux2-Klein 模型使用错误的 max_length=77 进行提示词标记化，而非 Hugging Face 参考实现中的 512。这导致长文本输入被不当截断，降低生成质量。根因是 Flux2-Klein 继承了 Flux1 的 text_encoder_extra_args 配置。

实现拆解 变更集中在文件 python/sglang/multimodal_gen/configs/pipeline_configs/flux.py：

- 在 Flux2PipelineConfig 类中新增 text_encoder_extra_args 字段，设置 max_length=512 等参数。
- 在 Flux2KleinPipelineConfig 的 tokenize_prompt 方法中，移除 tok_kwargs 中的 max_length，并硬编码为 512，确保使用正确长度。

评论区精华

- 测试文件名拼写错误：gemini-code-assist[bot] 指出新测试文件名 "klien" 拼写错误，应为 "klein"，作者 adityavaid 迅速修正。
- 代码行必要性讨论：mickqian 建议移除 tok_kwargs.pop("max_length", None) 行，但 adityavaid 解释该操作是必要的，以防止向 tokenizer 传递多个 max_length 值导致错误。

风险与影响 风险较低：变更仅影响 Flux2-Klein 模型的标记化逻辑，修复了已知 bug。添加的回归测试减少了未来回归风险。影响范围限于使用该模型的用户，提升长提示处理准确性。

关联脉络 本 PR 与 Issue #21372 直接相关，解决了扩散模型配置继承问题。历史 PR 如 #20633 和 #20706 也涉及扩散模型管道配置的优化，显示团队在持续改进该模块的健壮性和一致性。