

PR #32639 完整报告

sgl-project/sglang

docs: clarify diffusion stage reuse guidance

合并时间: 2026-07-28 19:52

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/32639>

执行摘要

本次 PR 为纯文档变更, 优化了 SGLang-Diffusion 新模型支持文档中的 stage 复用指南。通过按优先级重新排列 stage 选择顺序 (native stages → native-stage subclass → custom single-purpose stage → aggregated custom stage), 并更新 pipeline 接线示例, 使模型贡献者能更快选择正确的开发路径, 降低 review 负担。

功能与动机

PR 标题为 "docs: clarify diffusion stage reuse guidance", 动机源自保持新模型端口更小、审查更快、更贴近共享优化运行时路径的目标。文档原来将 aggregated custom stage 作为模型特定路径的默认形状, 现在明确将其降为最后手段, 并补充了更细的阶梯分类。

实现拆解

1. 调整 stage 形状优先级表格: 在 `support_new_models.mdx` 中, 将 "Standard stages" 和 "Model-specific pre-processing" 替换为更细致的四级分级: Native stages、Native-stage subclass、Custom single-purpose stage、Aggregated custom stage, 每级给出明确的适用场景和布局示例。
2. 更新 pipeline 接线示例: 重新编排 `MyModelPipeline.create_pipeline_stages` 的示例代码, 示范如何优先选择 native-stage subclass (例如 `TextEncodingStage` 的子类) 而不是直接创建自定义 stage, 以及如何将一个窄用途的 custom stage 插入到 native stage 序列中而非整体替换。
3. 标记聚合 stage 为最后手段: 在文档中强调只有当多个准备步骤确实不可拆分时才使用 `BeforeDenoisingStage` 这种聚合形状, 避免将其当作默认方案。
4. 补充自动化测试指引: 在文档末尾增加 CI 测试阶段的说明, 帮助贡献者确认 stage 配置正确性。

无可用源码片段 (本次变更为纯文档修订)。

评论区精华

无实质 review 讨论, 仅包含 Gemini Code Assist 自动评论 (已失效)。

风险与影响

- 风险: 文档变更无运行时风险, 但若指导不准确可能引导贡献者选择不恰当的 stage 形状。本次变更通过明确分级降低了此风险。

- 影响：主要影响后续添加 diffusion 模型的开发者，帮助其更快做出设计决策；对现有模型无影响。

关联脉络

该 PR 与近期 diffusion 相关 PR（如 #32420 修复 offload stride 丢失）无直接代码关联，但属于持续优化 diffusion 模型支持体系的一部分，与 #32542（Kimi-K3 cookbook）共同完善新模型 onboarding 体验。