PR #23553 完整报告

sgl-project/sglang

[DOC] Add DFLASH speculative decoding documentation

合并时间：2026-04-25 08:18

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23553

执行摘要

本 PR 为 speculative decoding 文档新增 DFLASH Decoding 完整章节，包括参数表、使用示例、决策指南和对比表格。最终仅修改 docs_new/ 下的一个文件。Review 过程中完善了约束条件和参数命名。

功能与动机

DFLASH是SGLang支持的另一种投机解码算法，使用专用的草稿模型检查点进行线性块验证。需要官方文档指导用户配置和使用。本 PR 填补了这一空白。

实现拆解

1. 添加 DFLASH 章节：在 docs_new/docs/advanced_features/speculative_decoding.mdx 末尾新增 ## DFlash Decoding 小节，包含算法简介、参数表格（--speculative-dflash-block-size、--speculative-dflash-draft-window-size 等）和 Python 调用示例。
2. 更新决策指南：在文档开头的决策列表中增加一行，提示“你有 DFlash 草稿检查点：使用 DFLASH 算法”，并列出所需参数。
3. 更新对比表格：在快速对比表格中新增 DFLASH 行，列出约束条件：No --enable-dp-attention、pp_size == 1、禁用 overlap scheduler 和 mixed chunked prefill。
4. Review 修正：根据 gemini-code-assist[bot] 和 zijiexia 的建议，在约束条件中补充 mixed chunked prefill 禁用说明，并将参数 block_size 改为具体参数名 --speculative-num-draft-tokens。
5. 文档目录清理：根据 b8zhong 的建议，放弃对旧 docs/ 目录的修改，仅保留 docs_new/ 下的更新（对应 commit "Drop legacy speculative decoding docs update"）。

以下是 DFLASH 参数表格的关键片段（摘自最终版本）： <!--DFLASH参数表格：每个参数的具体用途和默认值--> <tdstyle={{padding:"9px 12px",fontWeight:500,backgroundColor:"rgba(255,255,255,0.05)"}}><code>--speculative-dflash-block-size</code></td> <tdstyle={{padding:"9px 12px",backgroundColor:"rgba(255,255,255,0.05)"}}>DFlash块大小（标记为`N`）。草稿模型一次预测N个token。</td> <tdstyle={{padding:"9px 12px",backgroundColor:"rgba(255,255,255,0.05)"}}><code>5</code></td> <!--窗口大小参数，与block-size关系约束--> <tdstyle={{padding:"9px 12px",fontWeight:500,backgroundColor:"rgba(255,255,255,0.05)"}}><code>--speculative

-dflash-draft-window-size</code></td> <tdstyle={{padding:"9px 12px",backgroundColor:"rgba(255,255,255,0.05)"}}>草稿KV滑动窗口大小。若设置则必须>=<code>speculative-num-draft-tokens</code>。</td> <tdstyle={{padding:"9px 12px",backgroundColor:"rgba(255,255,255,0.05)"}}><code>None</code></td>

评论区精华

- gemini-code-assist[bot]和 zijiexia强调在表格中明确 mixed chunked prefill 被禁用，并使用完整参数名："SGLang will automatically disable --enable-mixed-chunk."
- gemini-code-assist[bot]建议参数描述中使用 --speculative-num-draft-tokens 取代 block_size，增强可读性。
- b8zhong建议仅修改 docs_new 文件夹，避免两个文档目录不同步。

风险与影响

风险：参数描述与 CLI 实现不一致的风险。已通过 code review 确认示例命令与 server_args.py 匹配，约束条件准确，风险较低。 影响：用户可快速上手 DFLASH 解码，降低配置错误；团队减少重复答疑。无系统性能或安全影响。

关联脉络

本 PR 与同期文档 PR（如 #23684、#23622）都属于完善 SGLang 功能文档的工作。DFLASH 是近期引入的投机解码算法，其代码实现不在此 PR 范围内，后续可能有对应的代码文档更新。此外，历史 PR #21985 涉及注意力性能优化，间接提升投机解码效率。