

PR #32661 完整报告

sgl-project/sglang

docs(kimi-k3): clarify VLM compatibility

合并时间: 2026-07-29 11:53

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/32661>

执行摘要

此 PR 在 Kimi-K3 部署文档中添加了 VLM 兼容性表格，明确 PD、EPD、MM Encoder DP、CUDA IPC 和 ViT BCG 的支持状态和配置要点，帮助用户避免错误配置。

功能与动机

用户对 Kimi-K3 在 PD/EPD 部署下的 VLM 兼容性存在疑问，常误设

`--mm-enable-dp-encoder` 等参数。PR 通过清晰的表格公告各项特性的支持情况，并指明正确配置。

实现拆解

1. 在 Kimi-K3.mdx 的 Should ViT BCG be enabled? 小节前插入一个包含 5 行的 Markdown 表格。
2. 表格列包括 PD（支持）、EPD（支持，需 `--encoder-only/--language-only` 角色）、MM Encoder DP（内置，无需设置 `--mm-enable-dp-encoder`）、CUDA IPC（兼容单节点统一 /PD-prefill 角色，不替代 EPD 或 PD 传输）、ViT BCG（兼容统一 /encoder-only 角色，但建议仅在形状重复且有富余 HBM 时启用）。
3. 文档校验通过 pre-commit 和 git diff check。

无，纯文档变更。

评论区精华

无。

风险与影响

风险极低；若未来 Kimi-K3 实现与文档描述不一致，可能误导用户。影响限于阅读该文档的部署者。

关联脉络

该 PR 是 Kimi-K3 部署文档系列更新的延续，与其前序 PR #32654（更新 B300 配置）和 #32592（更新 GB200 性能数据）共同完善了 K3 部署指南。