PR #7048 完整报告

PaddlePaddle/FastDeploy

[Refactor] Replace --skip-mm-profiling with --deploy-modality text

合并时间：2026-03-30 10:40

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7048

执行摘要

- 一句话：重构多模态 token profiling 参数，用 deploy-modality text 替代 skip-mm-profiling，简化部署配置。
- 推荐动作：建议快速浏览此 PR，以了解如何复用现有参数简化多模态部署配置。重点关注 get_max_chunk_tokens 方法的逻辑调整，作为参数整合的设计示例。

功能与动机

原 --skip-mm-profiling 参数与 deploy-modality 参数功能存在语义重叠：当以纯文本模式部署时，本就不需要为多模态 token 预留显存。引入独立参数增加了配置复杂度，复用 deploy_modality 更加直观和一致（引用 PR body）。

实现拆解

主要修改文件为 fastdeploy/config.py。在 get_max_chunk_tokens 方法中添加条件检查 self.deploy_modality != DeployModality.TEXT，确保部署模式为 text 时跳过 mm token 叠加。PR body 中提到删除 EngineArgs.skip_mm_profiling 字段及相关启动参数，但文件列表未显示这些变更，可能已在其他提交中处理。

关键文件：

- fastdeploy/config.py（模块 Config）：核心变更点，修改了 get_max_chunk_tokens 方法以整合 skip_mm_profiling 功能到 deploy_modality 参数中

关键符号：get_max_chunk_tokens

评论区精华

Review 过程中没有实质性技术讨论，仅有 reviewer 'yuanlehome' 批准合并。Issue 评论涉及 cherry-pick 操作，未包含设计或实现细节。

- 暂无高价值评论线程

风险与影响

- 风险：回归风险：如果 deploy_modality 为 text 时逻辑不正确，可能导致显存预留不足或过剩。兼容性风险：用户需要迁移命令行参数，原 --skip-mm-profiling 不再可用，可能影响现有部署脚本。测试覆盖：PR body 指出已有单元测试覆盖，但新参数集成可能需额外验证。

- 影响：对用户影响：需更新部署命令，使用 --deploy-modality text 替代 --skip-mm-profiling，影响范围限于多模态部署场景。对系统影响：减少参数数量，简化配置管理，提升代码一致性。对团队影响：此重构体现了参数设计优化，可作为类似冗余消除的参考。
- 风险标记：配置变更迁移风险，核心逻辑变更

关联脉络

- PR #7068 Cherry-pick from PR 7048: 关联 cherry-pick PR，将重构应用到 release/2.4 和 release/2.5 分支，确保变更同步