# PR #25771 完整报告

sgl-project/sglang

fix(dsv4): drop stale pp_size=1 guard for V4 PD disaggregation

合并时间：2026-05-20 10:57

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25771

## 执行摘要

- 一句话：移除 V4 PD 与 PP 的冲突断言
- 推荐动作：值得快速合并，无争议的回归修复。可顺带关注：这是一个典型的“语义过期”bug——代码不做同步维护就变成 dead code 并引发误拦。团队未来可以在重构时及时标记或删除关联 guard，避免类似回退。

## 功能与动机

关联 Issue #25765 报告：在 rebase 后的版本中，尝试以 pp_size=8 启动 DeepSeek V4 PD 模式会因 early guard 导致 ValueError，阻止了原本应被支持的 PP 配置。PR body 明确引用该 issue 作为修复目标，说明这是 CI 回归引发的误拦。

## 实现拆解

1. 在 python/sglang/srt/arg_groups/deepseek_v4_hook.py 的 apply_deepseek_v4_defaults 函数中，找到第 54-59 行（即 if server_args.disaggregation_mode != "null" and server_args.pp_size > 1: raise ValueError(...) 的 guard 块）。
2. 直接移除这 7 行代码（一个 if 判断、一条注释、一个 raise 语句）。
3. 该函数剩余的所有校验（EAGLE 规范、SWA ratio、KV cache dtype 等）保持不变，且没有新增任何逻辑。
4. 没有任何测试、配置或部署文件配套变动。

关键文件：

- python/sglang/srt/arg_groups/deepseek_v4_hook.py（模块 启动校验；类别 source；类型 core-logic）：唯一修改的文件。删除了 apply_deepseek_v4_defaults 中一行 stale guard，该 guard 在 PR #24704 重构底层 KV 指针模块后已失效。

关键符号：apply_deepseek_v4_defaults

## 评论区精华

变更非常微薄且清晰：核心 reviewer ShangmingCai 给出 "LGTM" 并批准；gemini-code-assist[bot] 的自动 review 也未提出反对。没有实质性的技术讨论。

- 暂无高价值评论线程

## 风险与影响

- 风险：风险极低：移除的 guard 在 PR #23882 时已失效——PR #24704 已重构底层的 get_mla_kv_ptrs_with_pp 和相关 KV 传输逻辑，使其能够正确处理 V4 的 buffer-type-organized flat KV 指针。唯一可预见的风险是：如果后续有其他未记录的不兼容假设也依赖于该 guard，则移除后可能暴露出来，但根据上下文，这不太可能。
- 影响：影响范围较小：仅影响 DeepSeek V4 模型在 PD disaggregation 模式下同时启用 Pipeline Parallelism（pp_size > 1）的启动路径。修复后，原本被拦的用户（如 Issue 中 pp_size=8 配置）可以正常启动。对非 V4 模型、非 PD 模式、或 V4 PD 但 pp_size=1 的场景无影响。
- 风险标记：核心路径变更

## 关联脉络

- PR #24704 Add PP support for V4 PD disaggregation: 此 PR 重构底层 KV 指针传输模块，使 V4 PD 兼容 PP；但未清理本 PR 所移除的 stale guard，导致了后续回归。
- PR #23882 Initial V4 components ( 包括该 guard 的原始引入 ): 该 PR 首次加入 pp_size==1 的校验，原因是对应底层模块尚不支持 PP。