# PR #35786 完整报告

sgl-project/sglang

[docs] Retune the Qwen3.8-27B RTX 5090 DFLASH2 cells against 1cf2b8c

合并时间：2026-08-21 12:54

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/35786

## 执行摘要

- 一句话：按 1cf2b8c 重调 5090 DFLASH2 显存 pin 并置灰不可达项
- 推荐动作：值得精读，重点不在代码而在 PR body 的实测方法论：先确定用户实际被指示安装的构建，再逐个 pin 复测并记录 pool / slot / accept 数据，最后用 S=5 / S=4 槽位分析解释“为什么 fp32 低延迟不可达”。这种“不可达就置灰而不是硬给 pin”的取舍，以及“文档错误背后往往藏着产品 bug（#33352）”的根因意识，是这份文档 PR 最有借鉴价值的部分。对 sglang 维护者，建议单独跟进作者提出的 CUDA-graph 捕获失败诊断改进。

## 功能与动机

PR body 指出 #35663 添加的 DFLASH2 格子携带的 RTX 5090 pin 是在“页面从未告诉用户去安装”的构建上测量的；按用户实际被指示的构建（lmsysorg/sglang:qwen38-27b 容器 / 最新 PyPI 发布）复测时，DFLASH2 的 0.90 bf16 pin 在第一个请求即 OOM，0.945 fp32 pin 在 CUDA 图捕获阶段 OOM，而 DSPARK 全部通过。因此需要将 DFLASH2 的 pin 对准文档指定的 commit 1cf2b8c 重新标定，并把确实不可达的 fp32 Low-Latency 组合显式置灰而不是给出必失败的命令。

## 实现拆解

变更入口是 docs/cookbook/autoregressive/Qwen/Qwen3.8-27B.mdx 的安装引导与 docs/src/snippets/configs/Qwen/qwen3.8-27b.jsx 的命令生成器配置，按以下步骤拆解：

1. 固定构建 commit（mdx）：pip 与 Docker 两条安装路径均从“跟踪移动分支”改为显式 git checkout 1cf2b8c54d81802abc15dcf23a29b9cc687bc01e（对应 PR #35496），因为新 pin 是 commit 特定的；同时重写 DFLASH2 警告块，说明缺少 #35496 会在 NVFP4 格子启动时报 requires a dense FP16/BF16/FP32 target lm_head。
2. 重标 bf16 pin（jsx）：DFLASH2 的 RTX 5090 bf16 mem-fraction 由 0.90（沿用 DSPARK 的 pin）改为 0.88，原因是在 1cf2b8c 上 0.90 首个请求即 OOM；新 pin 下 pool 27,452 / 29,276，约 7.01 ms，accept 3.01。
3. 重标 fp32 pin（jsx）：fp32 由 0.945 + ratio 10 改为 0.895 + ratio 10。关键分析是：这些格子固定 --max-running-requests 1，balanced 的 mamba ratio 会为从不使用的并发预配 KV，反而让 state pool 缺一个 fp32 slot；只有 High-Throughput（S=4）可容纳，Low-Latency（S=5）无解。

4. 置灰 fp32 Low-Latency（jsx）：在 ssmDtype 的 float32 选项上对 32 GB RTX 5090 + DFLASH2 组合做 disabled，通过既有 reseatHiddenPicks 机制将用户选择自动移到 bf16；stripPrefixes 相应在 fp32 下同时剥离 --mem-fraction-static 与 --mamba-full-memory-ratio。
5. 验证配套：无自动化测试（PR 声明 docs only），验证依赖 PR body 中的 RTX 5090 32GB 实测表格（NVFP4 checkpoint、ISL 8192 / OSL 1024、concurrency 1、exact-shape checks），并对 fp32 Low-Latency 给出了从 0.86 到 0.945 的全区间失败数据。

关键文件：

- docs/src/snippets/configs/Qwen/qwen3.8-27b.jsx（模块 文档脚本；类别 source；类型 core-logic；符号 config.dflash, config.ssmDtype）：命令生成器的核心配置：重调 RTX 5090 下 DFLASH2 的 mem-fraction 与 mamba ratio pin，并通过 disabled 与 stripPrefixes 组合将 fp32 Low-Latency 置灰。
- docs/cookbook/autoregressive/Qwen/Qwen3.8-27B.mdx（模块 部署文档；类别 docs；类型 documentation）：用户实际看到的部署文档：pip 与 Docker 安装路径均改为固定 checkout 1cf2b8c，并更新 DFLASH2 构建警告与 Configuration Tips，确保用户使用的构建与 pin 的测量基准一致。

关键符号：config.dflash.flags, config.dflash.stripPrefixes, config.ssmDtype.options

## 关键源码片段

### docs/src/snippets/configs/Qwen/qwen3.8-27b.jsx

命令生成器的核心配置：重调 RTX 5090 下 DFLASH2 的 mem-fraction 与 mamba ratio pin，并通过 disabled 与 stripPrefixes 组合将 fp32 Low-Latency 置灰。

```
// 注：以下配置来自 docs/src/snippets/configs/Qwen/qwen3.8-27b.jsx,
// DFLASH2 格子的 RTX 5090 专属 pin 全部按 commit 1cf2b8c（PR #35496）复测。
{
  id: "dflash", label: "DFLASH2",
  // Trained block-diffusion draft, a separate checkpoint. The selector
  // projects through the target lm_head — including quantized heads — so
  // it runs on the NVFP4 checkpoint too.
  disabled: (sel) => sel.hw === "rtx5090" && sel.quant !== "nvfp4",
  disableReason:
    "On the 32GB RTX 5090 the DFlash2 draft model only fits on top of the NVFP4 weights",
  // fp32 是唯一需要覆盖 balanced ratio 的场景，因此该组合要把两个
  // 前缀同时剥离，再在 flags 里重新输出。
  stripPrefixes: (sel) =>
    sel.hw === "rtx5090"
      ? sel.ssmDtype === "float32"
        ? ["--mem-fraction-static", "--mamba-full-memory-ratio"]
        : ["--mem-fraction-static"]
      : [],
  flags: (sel) => [
    "--speculative-algorithm DFLASH",
```

```
      "--speculative-draft-model-path incoai/Qwen3.8-27B-DFlash2",
      "--speculative-num-draft-tokens 8",
      // 关键：bf16 用 0.88（0.90 在 1cf2b8c 上首个请求即 OOM）；fp32 用 0.895
      // 且必须配合 ratio 10，因为格子固定 --max-running-requests 1，balanced
      // 的 ratio 会把 KV 预配给从不使用的并发，反而饿死 state pool。
      ...(sel.hw === "rtx5090"
        ? sel.ssmDtype === "float32"
          ? ["--mem-fraction-static 0.895", "--mamba-full-memory-ratio 10"]
          : ["--mem-fraction-static 0.88"]
        : []),
    ],
  }
```

### docs/cookbook/autoregressive/Qwen/Qwen3.8-27B.mdx

用户实际看到的部署文档：pip 与 Docker 安装路径均改为固定 checkout 1cf2b8c，并更新 DFLASH2 构建警告与 Configuration Tips，确保用户使用的构建与 pin 的测量基准一致。

```
# For the DFLASH2 cells only — DFlash2 selector support is newer than the
# latest release, so build from the commit those cells were validated on
# instead of the line above:
git clone https://github.com/sgl-project/sglang.git && cd sglang
# 固定 checkout 1cf2b8c（对应 PR #35496），文档里所有 DFLASH2 pin 均以此 commit 为基准
# 之前"跟踪 main"的做法会让 pin 与用户实际构建错位，这正是本 PR 要消除的偏差
# shellcheck disable=SC2046
git checkout 1cf2b8c54d81802abc15dcf23a29b9cc687bc01e
uv pip install -e "python[all]"
```

## 评论区精华

本 PR 没有实质性的 review 讨论线程：zijiexia 直接 APPROVED，唯一评论是 Mintlify 机器人贴出的文档预览链接，不涉及技术决策。核心讨论沉淀在 PR body 的"Note for maintainers"里，作者给出了一段值得关注的分析：

> The underlying cause of the DFLASH2 retune is #33352, which removed the 4 GiB free-memory gate on automatic prefill CUDA-graph capture that #31204 added. Builds carrying that gate fall back to eager prefill when headroom is short; builds without it attempt the capture and fail with num_active_captures_ > 0 INTERNAL ASSERT FAILED ... please report a bug to PyTorch, which gives no hint that --mem-fraction-static is the knob.

即真正的根因在产品侧（#33352 移除显存余量门槛），文档只是暴露了症状；作者建议独立于本 PR 改善该报错的诊断信息，这一点未被任何维护者回复或行动跟进。

- 暂无高价值评论线程

## 风险与影响

- 风险：

1. 固定 commit 的时效性风险：1cf2b8c 会随 main 演进过时，用户按文档构建可能遇到历史 commit 与当前依赖（PyPI 包、CUDA 版本）的兼容问题；若 DFlash2 或量化路径后续有修复，文档不会自动跟进，这与 #35767 改用滚动 dev tag 的做法形成张力。
2. pin 值依赖实测环境：0.88 / 0.895 / ratio 10 是在 RadixArk/Qwen3.8-27B-NVFP4 checkpoint 与特定引擎内存分配行为下测得的；checkpoint 更新或 PyTorch 图捕获策略变化都可能让这些值再次 OOM。
3. 置灰逻辑回归风险：jsx 中 disabled / stripPrefixes / reseatHiddenPicks 的组合依赖选择器框架行为，改动无自动化测试，后续若有人整理该 snippet 结构可能破坏置灰与 pin 输出的联动。
4. 构建成本变化：Docker 路径从 docker pull 改为从源码 docker build，用户需要完整构建工具链和时间，文档未提示镜像构建耗时或失败排查入口。
5. 根因未修复：#33352 导致的 CUDA-graph 捕获无友好诊断问题仍存在于产品中，其他模型和显卡组合仍可能撞上同类无提示 OOM。 - 影响：直接影响面是 RTX 5090 32GB 上运行 Qwen3.8-27B NVFP4 + DFLASH2 的用户：此前按文档配置会 OOM 或图捕获失败，现在 bf16 与 fp32 High-Throughput 可获得实测可启动的配置，fp32 Low-Latency 不再给出必失败的命令，而会被自动切换到 bf16。对文档团队的影响是确立了“文档 pin 必须对准文档告诉用户的构建”的实证纪律，并示范了“不可达组合显式置灰”的处理方式。影响范围严格限定在 Qwen3.8-27B cookbook 页面与命令生成器配置，不修改任何运行时代码，因此对线上服务无影响。 - 风险标记：文档 pin 依赖固定 commit，无自动化测试，置灰逻辑依赖选择器框架，根因 #33352 诊断缺失

## 关联脉络

- PR #35753 [docs] Tell Qwen3.8-27B DFLASH2 users to build from main: 同一文档页面的直接前序：该 PR 指示 DFLASH2 用户从 main 构建；本 PR 进一步收敛为固定 checkout 1cf2b8c，并依赖其引入的按选项状态的验证徽章来支撑置灰逻辑。
- PR #35767 [docs] Point the Qwen3.8-27B DFLASH2 note back at the rolling dev image tag: 同一“DFLASH2 用什么构建”决策点的反复：该 PR 把 note 指回滚动 dev 镜像 tag，本 PR 则明确放弃滚动 tag、改为固定 commit 构建并重测所有 pin。
- PR #34247 [Docs] Standardize diffusion cookbook model pages: 建立了 docs/src/snippets/configs 下 JSX 命令生成器与 cookbook 校验脚本的基础设施，本 PR 的 qwen3.8-27b.jsx 修改正是运行在该文档工程框架之上。