

# PR #37301 完整报告

sgl-project/sglang

[Cookbook] Enable DSpark on the DeepSeek-V4 Flash Vision low-latency recipes

合并时间: 2026-09-01 07:43

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/37301>

## 执行摘要

本 PR 是 DeepSeek-V4 cookbook 的文档配套变更: 在 Flash Vision (Exp) 低延迟配方中启用 DSpark 投机解码。核心改动是在 `deepseek-v4.jsx` 的 5 个低延迟 recipe (B200 / B300 / GB200 / GB300 / H200) 中加入 `--speculative-algorithm DSPARK`, 移除 Playground 中 Flash Vision 变体的 DSpark 置灰规则, 并把 B200 的 MMMU-Pro 实测分数更新为 75.14% (开启 DSpark 后复测, 1730/1730 样例)。balanced / high-throughput 配方因依赖 DP Attention (与 DSpark 不兼容) 保持 target-only。纯文档变更, 无运行时代码改动。

## 功能与动机

PR body 明确动机: #37293 以 target-only 方式发布 Flash Vision cookbook 配方, 是因为 DSpark draft head 尚未在图像批次上验证; #37253 已把 DSpark 相关 hook (`lm_head` 委托、`set_dspark_layers_to_capture`) 接入 VL wrapper, 剩下的工作是在 B200 上完成验证并更新文档。

验证结果: MMMU-Pro (standard, 10-option) 75.14%, 1730/1730 样例全部完成, stop\_rate 100%, 0 截断、0 报错, decode accept 长度约 2.5–3.0。DSpark 与 DP Attention 不兼容是 balanced / high-throughput 保持 target-only 的根本原因。

## 实现拆解

- 解除 DSpark 选项的 Flash Vision 禁用: 在 `docs/src/snippets/configs/deepseek-ai/deepseek-v4.jsx` 的 `speculative.options` 数组中, `dspark` 选项的 `disable` 列表移除了针对 `variant: ["flash-vision"]` 的规则; 保留 `dpAttnOn: [true]` 与 `hw: ["mi300x", "mi355x"]` 两条不兼容约束。
- 低延迟 recipe 注入 DSPARK 参数: 在 B200 / B300 / GB200 / GB300 / H200 五个 flash-vision + fp4 + low-latency 单元格的 `flags` 中插入 `--speculative-algorithm DSPARK`, 位置紧接 `--tp 4`, 与 Flash Official (0731 / 0813) 标志顺序一致; 不追加 EAGLE 系列参数, 因为草稿头内置于主检查点。
- 基准数据更新: `deepseek-v4-benchmarks.jsx` 把 B200 Flash Vision 低延迟行的 `mmmu_pro_pct` 从 74.96 更新为 75.14, `notes` 补充 DSpark 与验证命令细节。
- 同步 MDX 叙述: `DeepSeek-V4.mdx` 两处关于 Flash Vision 投机解码的描述改为 "低延迟配方启用 DSpark (B200 已验证, 其余硬件 pending)", 并说明 balanced / high-throughput 因 DP Attention 保持 target-only, 且不传 EAGLE 三个参数 (`--speculative-num-steps`、`--speculative-eagle-topk`、`--speculative-num-draft-tokens`)。

5. 状态标注与测试配套：B200 低延迟行保持 verified: true 并保留未发布警告（依赖 #37253）；B300 / GB200 / GB300 / H200 低延迟行保持 verificationStatus: "in-progress"。纯文档变更，无测试配套（PR body 自述 N/A）。

### docs/src/snippets/configs/deepseek-ai/deepseek-v4.jsx

cookbook 配置面板的核心数据源。本 PR 在这里解除 Flash Vision 变体的 DSpark 禁用规则，并为 B200 / B300 / GB200 / GB300 / H200 五个低延迟 recipe 注入 --speculative-algorithm DSPARK，是启用 DSpark 的主要载体。

```
// ----- Card 4: "Speculative Decoding" -----
// dspark 选项: hide 只作用于普通 flash / pro 变体,
// Flash Official 与 Flash Vision 均允许选择 DSpark
speculative: {
  options: [
    { id: "current", label: "Inherited from base" },
    { id: "off", label: "Off (greedy)" },
    { id: "mtp-314", label: "EAGLE / MTP 3-1-4",
      flags: ["--speculative-algorithm EAGLE", "--speculative-num-steps 3",
        "--speculative-eagle-topk 1", "--speculative-num-draft-tokens 4"],
      hide: { variant: ["flash-official", "flash-vision", "pro-official"] } },
    { id: "dspark", label: "DSpark",
      // 草稿头随主检查点发布，因此只需一个参数即可启用
      flags: ["--speculative-algorithm DSPARK"],
      hide: { variant: ["flash", "pro"] },
      disable: [
        // Flash Vision 的禁用规则已移除（B200 上完成图像批次验证）；
        // 仅保留两条硬性不兼容约束：DP Attention 与 ROCm
        { when: { dpAttnOn: [true] },
          reason: "DSpark is not compatible with DP Attention on the current release." },
        { when: { hw: ["mi300x", "mi355x"] },
          reason: "DSpark currently requires CUDA." },
      ] },
  ],
},

// B200 + FP4 — Flash Vision (Exp) 低延迟 recipe
// verified: true 表示用下方精确命令完成过完整验证
{
  match: { hw: "b200", variant: "flash-vision", quant: "fp4", strategy: "low-latency", nodes:
    "single" },
  verified: true,
  warn: "DeepSeek-V4-Flash-Vision-Exp support has not shipped in an SGLang release yet
    (sglang PR 37253): Docker mode already points at the preview image; for Python mode install
    SGLang from that PR. See [Flash Vision notes](#vision-note).",
  env: [],
  flags: [
    "--model-path {{MODEL_NAME}}",
    "--tp 4",
    // 与 Flash Official 标志顺序一致，紧接 --tp 4;
```

```
// 不传任何 EAGLE 系列参数, draft 已内置在主检查点
"--speculative-algorithm DSPARK",
"--mem-fraction-static 0.85",
"--host {{HOST_IP}}",
"--port {{PORT}}",
],
},
```

## docs/src/snippets/configs/deepseek-ai/deepseek-v4-benchmarks.jsx

基准数据页。将 B200 Flash Vision 低延迟行的 MMMU-Pro 从 74.96 更新为 75.14, 并在 notes 中注明该分数是在开启 DSpark 后测得, 保证数据与配方一致。

```
{
  match: { hw: "b200", variant: "flash-vision", quant: "fp4", strategy: "low-latency", nodes:
    "single" },
  slang_version: "dev-dsv4-flash-vision",
  // 开启 DSpark 后的 MMMU-Pro 实测值: 1730/1730 样例全部完成,
  // stop_rate 100%, 0 截断、0 报错; decode 接受长度约 2.5–3.0
  accuracy: { mmmu_pro_pct: 75.14 },
  notes: "MMMU-Pro (standard, 10-option) measured with sgl-eval on 4xB200 (TP=4) at
    temperature 1.0, top-p 0.95, --reasoning-effort max, with the bundled DSpark head enabled (--
    speculative-algorithm DSPARK).",
},
```

## 评论区精华

本 PR 没有实质性 review 讨论: [wisclmy0611](#) 直接 approve (空评论), 唯一的评论来自 mintlify[bot] 的预览部署通知。技术权衡全部沉淀在 PR body 中:

B200 stays **verified: true**, re-verified with the exact shipped command: MMMU-Pro (standard, 10-option) 75.14% — 1730/1730 examples, **stop\_rate** 100%, 0 truncations, 0 errors.

Balanced / high-throughput recipes stay target-only: they run DP Attention, which DSpark is incompatible with on current releases.

## 风险与影响

- 多硬件验证缺口: B300 / GB200 / GB300 / H200 的低延迟行已写入 DSPARK 参数但保持 verificationStatus: "in-progress", 文档配置先于这些平台的验证发布; 状态字段显式可见, 风险可控。
- 依赖未发布特性: Flash Vision 支持依赖 #37253, 尚未进入正式 release; warn 已提示 Python 模式需从该 PR 源码安装, 用户若用旧版本照抄命令会启动失败。
- 配置与运行时约束耦合: DSpark × DP Attention 的不兼容仅体现在配置面板 disable 规则中, 绕过 UI 手动组合参数时文档无法强制约束。
- 精度数据可比性: 75.14% 与先前 74.96% 的差异来自开启 DSpark 后的命令级复测; 读者需注意该分数带 DSpark, 而非 target-only 配置。

- 影响范围：仅 DeepSeek-V4 cookbook 页面（3 个文件、15 增 11 删），无运行时代码、无测试、无配置默认值变更。

## 关联脉络

- #37293是本 PR 的直接前身：首次把 Flash Vision 变体加入 DeepSeek-V4 页面，全部配方 target-only；本 PR 在其基础上完成低延迟配方的 DSpark 启用与 B200 验证。
- #37253（PR body 引用）为 DSpark 在 VL wrapper 中的接入 PR，是本次文档变更成立的运行时前提。
- 结合 0731 / 0813 检查点的既有做法（DSpark 内置于主检查点、不传 EAGLE 参数），可以看到 DeepSeek-V4 系列文档在 " 投机解码参数最小化 " 原则下的一致性演进：draft head 随主检查点发布，cookbook 只暴露 `--speculative-algorithm DSPARK` 一个开关。