PR #37293 完整报告

sgl-project/sglang

[Cookbook] Add DeepSeek-V4-Flash-Vision-Exp to the DeepSeek-V4 page

合并时间：2026-09-01 05:52

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/37293

1. 执行摘要

PR #37293 是一个纯文档站变更：在 DeepSeek-V4 cookbook 页面新增实验性多模态变体 Flash Vision（deepseek-ai/DeepSeek-V4-Flash-Vision-Exp）。改动分两层——配置数据层新增 flash-vision 变体、模型映射、部署格、MMMU-Pro 基准与投机解码门控；文档渲染引擎层把 Docker 镜像解析链扩展出 hw|variant|quant → variant|quant 两级键，使 Flash Vision 格自动命中 preview 镜像 lmsysorg/sglang:dev-dsv4-flash-vision，其余变体解析结果不变。B200 FP4 low-latency 格已端到端验证（MMMU-Pro 74.96%，4×B200 TP=4），其余格为 in-progress / pending。整个变更依赖引擎支持 PR #37253，release 前以 preview 镜像引导用户，并有配套的 cookbook-add-model 技能文档同步。

2. 功能与动机

PR body 明确动机：把 DeepSeek 首个实验性多模态 V4 checkpoint 接入既有页面，而不是新建页面——

> Add the experimental multimodal checkpoint deepseek-ai/DeepSeek-V4-Flash-Vision-Exp (engine support: #37253) to the DeepSeek-V4 cookbook as a new Flash Visionvariant on the existing page.

该 checkpoint 是 0731 Flash 基座 + vision encoder / aligner（305B 总参数），MDX 正文称其为 DeepSeek's first experimental multimodal V4 checkpoint。文档需要给出：已验证部署命令（B200 FP4 low-latency）、实测精度（MMMU-Pro 74.96%）、以及一组未验证不冒充已验证的状态治理——balanced / high-throughput 标记 verificationStatus: "in-progress"，其它硬件格 pending。同时由于 Flash Vision 与 Flash / Flash Official 共享 hw / quant / strategy 维度值，必须扩展镜像解析键位才能让变体级 preview 镜像生效。

3. 实现拆解

1. 配置数据扩展（deepseek-v4.jsx）：新增 flash-vision 变体（305B・Exp）与 "flash-vision|fp4" 模型映射；在 accuracyLabels 追加 mmmu_pro_pct；在 benchmarkCommands.accuracy 增加按变体键控的 MMMU-Pro 复现命令；dockerImages 增加 "flash-vision|fp4" 指向 preview 镜像；投机解码面板对 flash-vision 隐藏 EAGLE 预设、禁用 DSpark 并附原因。最后追加 B200 FP4 三策略格（low-latency 已验证，其余 in-progress）以及 B300 / GB200 / GB300 / H200 / H100 的 pending 格，所有格带 warn banner。

2. 基准数据（deepseek-v4-benchmarks.jsx）：新增 Flash Vision 基准条目，只有 B200 low-latency 格携带 sglang_version 与实测精度 74.96%（附测量条件），其余占位。注意 h100 只配 balanced 一格，与其它硬件三策略齐全不一致。
3. 共享引擎镜像解析链（_deployment.jsx 与 _playground.jsx）：两处同一逻辑——镜像选择从 hwlquantlstrategy → hwlquant → hw 扩展为 hwlvariantlquant → variantlquant → hwlquantlstrategy → hwlquant → hw。新键插在链首，历史配置（无 variant 键）解析结果逐字节不变。两个引擎必须同步改，因为 Mintlify 剥离共享模块状态，playground 只能复制一份解析逻辑。
4. 页面正文（DeepSeek-V4.mdx）：Deploy 面板上方加 <Note> 预告 preview 镜像；§1 变体表与资源行增加 Flash Vision；#vision-note 锚点章节说明 preview 构建、span 对齐 chunked prefill 与 radix-cache 行为、shared-experts fusion 自动关闭、target-only 投机与评测 max_tokens 提示；§3.4 补 DSpark 说明；§3.5 增加视觉 image-input 示例（输出待补）。
5. 配套技能文档（cookbook-add-model）：config.jsx.tmpl、SKILL.md、authoring-reference.md 同步五级镜像键契约，并重申询问作者实际构建、不要猜 release 标签。

测试与验证配套：无自动化测试文件；PR body 报告 mint validate 与 mint broken-links 通过，并在 mint dev 下手工冒烟 badge 状态、镜像解析、精度卡作用域、playground 门控与锚点链接。

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

本 PR 的核心配置数据文件：新增 flash-vision 变体、模型映射、preview 镜像键、MMMU-Pro 指标与复现命令、投机解码门控，以及 B200 / B300 / GB200 / GB300 / H200 / H100 的部署格。

```
// ===== Flash Vision (Exp) 相关配置摘录（deepseek-v4.jsx） =====
// 变体表：新增 305B 实验性多模态条目（0731 Flash 基座 + 视觉编码器）。
variants: [
  { id: "flash", label: "Flash", subtitle: "284B" },
  { id: "flash-official", label: "Flash Official", subtitle: "284B · 0731" },
  { id: "flash-vision", label: "Flash Vision", subtitle: "305B · Exp" },
  { id: "pro", label: "Pro", subtitle: "1.6T" },
  { id: "pro-official", label: "Pro Official", subtitle: "1.6T · 0813" },
],
// 模型映射：该变体只发布 FP4 量化，直接指向官方实验 checkpoint。
modelNames: {
  "flash-visionlfp4": "deepseek-ai/DeepSeek-V4-Flash-Vision-Exp",
},
// 专属 preview 镜像：引擎支持尚未进入 release（见 sgl-project/sglang#37253），
// 由镜像解析链保证只有 Flash Vision 格命中它；其余硬件键保持 `:latest`。
dockerImages: {
  "flash-visionlfp4": "lmsysorg/sglang:dev-dsv4-flash-vision",
},
// 投机解码门控：checkpoint 自带 DSpark draft head，但“图像批次 + 投机解码”
// 尚未验证，因此 EAGLE 预设对该变体隐藏、DSpark 禁用，配方一律 target-only。
```

```
  "mtp-314": { hide: { variant: ["flash-official", "flash-vision", "pro-official"] } },
  dspark: {
    disable: [
      { when: { variant: ["flash-vision"] },
        reason: "The Flash Vision checkpoint bundles a DSpark head, but speculative decoding is
        not yet verified with image inputs — the cookbook recipes run target-only for now." },
    ],
  },
  // B200 + FP4 low-latency 格：唯一端到端验证的部署形态（4×B200、TP=4）。
  {
    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).",
    flags: ["--model-path {{MODEL_NAME}}", "--tp 4", "--mem-fraction-static 0.85", "--host {{HOST_
    IP}}", "--port {{PORT}}"],
  },
```

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

基准数据入口：为 Flash Vision 各格新增基准条目，只有已实测的 B200 low-latency 格携带精度与 sglang 版本，其余占位待终验。

```
  // ===== B200 + FP4 — Flash Vision (Exp) 基准条目（deepseek-v4-benchmarks.jsx） =====
  // 只有 low-latency 格带实测数据：MMMU-Pro (standard, 10-option) 74.96%，
  // 测量条件与 sglang 构建版本一并记录，保证可复现；其余格先占位，待终验后翻转。
  {
    match: { hw: "b200", variant: "flash-vision", quant: "fp4", strategy: "low-latency", nodes:
    "single" },
    sglang_version: "dev-dsv4-flash-vision",
    accuracy: { mmmu_pro_pct: 74.96 },
    notes: "MMMU-Pro (standard, 10-option) measured with sgl-eval on 4×B200 (TP=4) at
    temperature 1.0, top-p 0.95, --reasoning-effort max.",
  },
  // balanced / high-throughput 与其它硬件（B300 / GB200 / GB300 / H200 / H100）
  // 均为 pending 占位；h100 目前只配 balanced 一格，与其他硬件三策略齐全不一致。
  { match: { hw: "b200", variant: "flash-vision", quant: "fp4", strategy: "balanced", nodes: "single" }
  },
  { match: { hw: "b200", variant: "flash-vision", quant: "fp4", strategy: "high-throughput", nodes:
  "single" } },
  { match: { hw: "h100", variant: "flash-vision", quant: "fp4", strategy: "balanced", nodes: "single" }
  },
```

docs/src/snippets/_deployment.jsx

共享部署引擎：Docker 镜像解析链扩展出变体级键位，是让 Flash Vision 命中 preview 镜像、同时保持其它 cookbook 向后兼容的关键逻辑改动。

```
  // ===== Docker 镜像解析链（_deployment.jsx；_playground.jsx 为同一逻辑的副本） =====
```

```
// 解析优先级从具体到宽泛：命中第一个存在的键即返回。
// 新增的 ``hw|variant|quant`` 与 ``variant|quant`` 两级，用于“同一硬件 / 量化下，
// 某个 checkpoint 变体需要专属镜像”的场景（如 Flash Vision 的 preview 构建）；
// 原有 ``hw|quant|strategy → hw|quant → hw`` 链放在其后，保证不带 variant 键的
// 历史配置解析结果与改动前完全一致。
const di = config.dockerImages || {};
const image = di[`${sel.hw}|${sel.variant}|${sel.quant}`] // 例：“b200|flash-vision|fp4”
  || di[`${sel.variant}|${sel.quant}`] // 例：“flash-vision|fp4”（跨硬件共享镜像）
  || di[`${sel.hw}|${sel.quant}|${sel.strategy}`] // 原有策略级键（如 spec-decoding preview 镜像）
  || di[`${sel.hw}|${sel.quant}`] // 原有“硬件 + 量化”键
  || di[sel.hw] // 原有硬件级键
  || "lmsysorg/sglang:dev"; // 兜底 dev 镜像
```

5. 评论区精华

本 PR 几乎没有形成 review 交锋：review_comments_count = 0，唯一 approval（wisclmy0611）为空正文，唯一 issue 评论是 mintlify[bot] 的预览部署通知。值得提炼的讨论点全部来自 PR body 的自述性设计权衡：

- 镜像解析链为何要动共享引擎：Flash Vision 与 Flash / Flash Official 共享 hw / quant / strategy 维度值，旧键位无法把镜像作用域切到变体级，因此新增两级变体键。
- 双引擎重复维护的架构约束：> Both engines carry the change because the playground duplicates image resolution by design (Mintlify strips shared module state). 这是 Mintlify 平台限制下的刻意取舍，不是冗余代码。
- 验证数据治理纪律：> recorded as per-cell accuracy on the verified low-latency cell only — the in-progress cells stay on the "pending" state. 只有实测格标 verified，其余用 verificationStatus 三态占位，避免未验证数据冒充已验证。

6. 风险与影响

- 共享引擎回归面：_deployment.jsx / _playground.jsx 被所有 cookbook 页面复用。新键插在解析链首，理论上对不带 variant 键的存量配置向后兼容；但若某现存配置恰好定义了 variant|quant 或 hw|variant|quant 键（此前为永不命中的死键），行为会变化，建议对其它 cookbook 页面做一次镜像解析冒烟回归。
- 全局指标行扩容：mmmu_pro_pct 加入 accuracyLabels 后，Flash / Pro 等没有该数据的变体，基准卡也会渲染该行，可能出现空值或 N/A，需要确认渲染层对缺失精度的处理。
- 数据不对称：h100 的 flash-vision 只有 balanced 一格（cells 与 benchmarks 均如此），缺少 low-latency / high-throughput 入口，用户按 h100 查找时入口不完整。
- 依赖未发布引擎：所有 Flash Vision 格依赖 #37253；Python 模式命令要求用户自行安装 PR 分支，若照抄到 release 环境会失败（warn banner 已缓解）。#37253 合入并发布后，需要把 preview 镜像统一替换回 :latest，这是明确的后续维护项。
- 本 PR 不含任何引擎运行时代码，不影响 SGLang 服务稳定性。

7. 关联脉络

本 PR 与引擎支持 PR #37253 构成“支持 + 指南”的配对关系：文档侧所有 Flash Vision 格与 preview 镜像都指向它，待其合入并发布后需做一轮镜像 tag 替换。测试侧，同仓库近期有 #37214 重新启用 DeepSeek-V4-Flash 的 NPU 夜间性能用例，说明该模型家族在测试与文档两条线上并行铺开。更宏观的演进信号在 .claude/skills/cookbook-add-model：dockerImages 契约从 hw → hwlquant → 五级键链持续扩展，意味着 cookbook 基础设施正从“一页一模型”走向“一页多变体、变体级差异化”，本次新增的 flash-vision 是其第一个变体级 preview 镜像用例。