

PR #37351 完整报告

sgl-project/sglang

[Cookbook] Add NVFP4 options for DeepSeek-V4 Flash Official (0731) and Pro Official (0813)

合并时间: 2026-09-01 16:03

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/37351>

执行摘要

本 PR 为 DeepSeek-V4 的官方变体 (Flash Official 0731 和 Pro Official 0813) 添加了 NVFP4 量化配置选项, 扩展了 cookbook 的硬件覆盖范围。变更集中在文档配置文件, 无核心引擎代码改动, 风险较低, 主要提升文档完整性和用户体验。

功能与动机

NVIDIA 发布了 DeepSeek-V4 官方变体的 NVFP4 量化检查点 ([nvidia/DeepSeek-V4-Flash-0731-NVFP4](#) 和 [nvidia/DeepSeek-V4-Pro-0813-NVFP4](#)), 但现有 cookbook 仅支持原始 Flash/Pro 变体的 NVFP4。本 PR 旨在让这些新检查点可在 B200、B300、GB200、GB300 等 Blackwell 硬件上一键部署。

实现拆解

1. 模型映射更新: 在 `deepseek-v4.jsx` 的 `modelNames` 对象中, 添加了 `flash-officialnvfp4` 和 `pro-officialnvfp4` 键, 映射到对应的 NVIDIA Hugging Face 仓库。
2. 命令行配置单元: 为四个硬件组合 (B200、B300、GB200、GB300) 新增低延迟策略配置, 关键差异是使用 `--speculative-algorithm DSPARK` (因为官方检查点捆绑了 DSpark 草稿头), 而非 EAGLE/MTP 标志。所有新单元标记为未验证 (`verified: false` 或 `verificationStatus: "in-progress"`)。
3. Docker 镜像兼容: 在 `dockerImages` 中, 为 NVFP4 场景指定 nightly 镜像 (如 `lmsysorg/sglang:dev`), 因为稳定版本 (`v0.5.18`) 加载这些检查点时会崩溃 (需 PR #36275 修复)。
4. 基准数据补充: 在 `deepseek-v4-benchmarks.jsx` 中, 为 B200 上的官方 NVFP4 配置添加了 GSM8K 和 AIME25 准确度数据, 但未提供速度数据。
5. 文档文本更新: 在 `DeepSeek-V4.mdx` 中, 扩展了 NVFP4 部分, 列出官方检查点链接, 并明确 DSpark 算法的使用场景。

[docs/src/snippets/configs/deepseek-ai/deepseek-v4.jsx](#)

主要配置文件, 包含模型映射、命令行配置单元和 Docker 镜像映射的更新, 是 PR 的核心变更。

以下片段展示了模型名称映射的更新和关键配置单元的新增:

```
// 模型名称映射: 为官方变体添加 NVFP4 路径
modelNames: {
```

```

"flashlfp4": "deepseek-ai/DeepSeek-V4-Flash",
"flashlfp8": "deepseek-ai/DeepSeek-V4-Flash",
"flashlnvfp4": "nvidia/DeepSeek-V4-Flash-NVFP4",
"flash-officiallfp4": "deepseek-ai/DeepSeek-V4-Flash-0731",
"flash-officiallnvfp4": "nvidia/DeepSeek-V4-Flash-0731-NVFP4", // 新增: 官方 Flash NVFP4
"flash-visionlfp4": "deepseek-ai/DeepSeek-V4-Flash-Vision-Exp",
"prolfp4": "deepseek-ai/DeepSeek-V4-Pro",
"prolfp8": "deepseek-ai/DeepSeek-V4-Pro",
"prolnvfp4": "nvidia/DeepSeek-V4-Pro-NVFP4",
"pro-officiallfp4": "deepseek-ai/DeepSeek-V4-Pro-0813",
"pro-officiallnvfp4": "nvidia/DeepSeek-V4-Pro-0813-NVFP4", // 新增: 官方 Pro NVFP4
// H200 FP8 需要特殊打包 (Hopper 不能运行 FP4 混合 Instruct)
"h200lflashlfp8": "sgl-project/DeepSeek-V4-Flash-FP8",
"h200lprolfp8": "sgl-project/DeepSeek-V4-Pro-FP8",
// AMD FP8 使用 sgl-project 打包
"mi300xlflashlfp8": "sgl-project/DeepSeek-V4-Flash-FP8",
"mi355xlflashlfp8": "sgl-project/DeepSeek-V4-Flash-FP8",
"mi355xlprolfp8": "sgl-project/DeepSeek-V4-Pro-FP8",
},

```

// Docker 镜像映射: NVFP4 检查点需要 nightly 镜像, 因为 v0.5.18 有崩溃问题

```

dockerImages: {
  // Flash Vision (Exp) 支持尚未发布, 需要预览构建
  "flash-visionlfp4": "lmsysorg/sglang:dev-dsv4-flash-vision",
  // NVFP4 检查点在 v0.5.18 上崩溃 (MXFP4 打包的 MTP 层的 FP8 委托需要 #36275 修复)
  "b200lnvfp4": "lmsysorg/sglang:dev",
  "b300lnvfp4": "lmsysorg/sglang:dev",
  "gb200lnvfp4": "lmsysorg/sglang:dev",
  "gb300lnvfp4": "lmsysorg/sglang:dev",
  h100: "lmsysorg/sglang:latest",
  // ... 其他硬件使用 latest
},

```

// 配置单元示例: B200 + NVFP4 — 官方变体使用 DSPARK 算法

```

{
  match: { hw: "b200", variant: "flash-official", quant: "nvfp4", strategy: "low-latency", nodes:
    "single" },
  verified: true, // 在 B200 上已验证
  env: [],
  flags: [
    "--trust-remote-code",
    "--model-path {{MODEL_NAME}}",
    "--tp 4",
    "--moe-runner-backend flashinfer_trtllm_routed",
    "--speculative-algorithm DSPARK", // 官方检查点捆绑 DSpark 草稿头
    "--disable-flashinfer-autotune",
    "--swa-full-tokens-ratio 0.1",
    "--host {{HOST_IP}}",
    "--port {{PORT}}",
  ],
}

```

```
],  
},
```

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

添加官方 NVFP4 变体的准确度基准数据，为用户提供性能参考。

以下片段展示了新增的准确度数据：

```
// B200 + NVFP4 — 官方变体准确度数据（无速度数据）  
{  
  match: { hw: "b200", variant: "flash-official", quant: "nvfp4", strategy: "low-latency", nodes:  
    "single" },  
  sglang_version: "dev@07c8f7294", // 基于开发版本  
  accuracy: { gsm8k_pct: 96.82, aime25_pct: 98.96 }, // 新增准确度指标  
},  
{  
  match: { hw: "b200", variant: "pro-official", quant: "nvfp4", strategy: "low-latency", nodes:  
    "single" },  
  sglang_version: "dev@07c8f7294",  
  accuracy: { gsm8k_pct: 96.44, aime25_pct: 98.33 },  
},
```