PR #26057 完整报告

sgl-project/sglang

[docs] DeepSeek-V4 cookbook: split Quantization axis, add H100 SGLang FP8

合并时间：2026-05-22 15:52

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26057

执行摘要

此 PR 重构 DeepSeek-V4 部署命令生成器的 UI，将量化选择从硬件标签中拆分为独立轴，并新增 H100 SGLang FP8 路径支持。本质是文档和交互式生成器的更新，不涉及引擎变更，风险低，但使部署配置结构更清晰。

功能与动机

原设计将 FP4/FP8 编码进硬件标签（如 H200 (FP8) vs H200 (FP4)），混淆了两个正交维度，且完全隐藏了 H100 SGLang FP8 路径。用户需要根据“硬件我有”和“想运行的检查点”进行心理翻译。本次分解使硬件行只代表 GPU 型号，量化作为独立选择，并补全了 H100 FP8 路径的基准数据。

实现拆解

1. UI 结构重构：在 docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx 的 options 对象中，将硬件行中的量化后缀移除，并新增独立的 quantization 轴（fp4 / fp8），同时移除原来独立的 h200-fp4 条目。
2. 内部兼容层：定义 effHw(hardware, quantization) 函数，将（硬件，量化）映射回旧内部标识（如 h200 + fp4 → h200-fp4，h100 + fp8 → h100-fp8），使得 HW_SIZE_SPEC、VERIFIED_RECIPES 等常量无需改动。
3. 约束与自动回退：新增 FP8_SUPPORTED_HARDWARE 集合，在非 Hopper 硬件上禁用 FP8；扩展 handleRadioChange 逻辑，当用户选择无效组合时自动回退（如 FP8 → FP4，Pro → Flash，cp/pd-disagg → low-latency）。
4. 新增 H100 FP8 规格：在 HW_SIZE_SPEC 中增加 h100-fp8lsmall 条目，并给出了已验证的 low-latency / balanced / max-throughput 配方。
5. MDX 文档同步：在 docs_new/cookbook/autoregressive/DeepSeek/DeepSeek-V4.mdx 中更新了表格、Hopper 说明，将示例输出包裹在 Accordion 内并替换为真实基准数据，重新组织第 5 节顺序（先准确率后速度），填入 GSM8K/MMLU/ 吞吐量数值。

docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx

核心变更：重构命令生成器 UI，分离量化选项，新增 effHw 映射函数和约束逻辑。

```
// 新增的量化选择轴，仅 Hopper 支持 FP8
const quantizationOptions = {
  name: "quantization",
  title: "Quantization",
```

```
    items: [
      { id: "fp4", label: "FP4", default: true },
      { id: "fp8", label: "FP8", default: false, subtitle: "H100/H200 only" },
    ],
  };

  // 将用户 (hardware, quantization) 映射为内部标识，保持旧常量兼容
  const effHw = (hardware, quantization) => {
    // H200 + fp8 → "h200" (FP8 路径直接映射到旧 h200)
    // H200 + fp4 → "h200-fp4" (Marlin FP4 路径 )
    if (hardware === "h200") return quantization === "fp8" ? "h200" : "h200-fp4";
    // H100 + fp8 → "h100-fp8" ( 全新标识 )
    // H100 + fp4 → "h100" (Marlin)
    if (hardware === "h100") return quantization === "fp8" ? "h100-fp8" : "h100";
    return hardware;
  };

  // Hopper GPUs 支持 SGLang FP8 重打包
  const FP8_SUPPORTED_HARDWARE = new Set(["h100", "h200"]);

  // Marlin (FP4) 路径不支持 cp 和 pd-disagg 配方
  const MARLIN_UNSUPPORTED_RECIPES = new Set(["cp", "pd-disagg"]);
  // 注意：原来使用 MARLIN_HARDWARE，现改为 MARLIN_EFFHW，使用 effHw 后的标识
  const MARLIN_EFFHW = new Set(["h200-fp4", "h100"]);
  const MARLIN_LABEL = { "h200-fp4": "H200 (FP4)", h100: "H100 (FP4)" };

  // MegaMoE 仅在 Blackwell + DeepEP 配方下可用，禁用集合简化为只含 "h100" 和 "h200"
  const MEGAMOE_UNSUPPORTED_RECIPES = new Set(["low-latency", "cp"]);
  const MEGAMOE_UNSUPPORTED_HARDWARE = new Set(["h100", "h200"]);
  const isMegamoeUnsupported = (vals) =>
    MEGAMOE_UNSUPPORTED_HARDWARE.has(vals.hardware) ||
    MEGAMOE_UNSUPPORTED_RECIPES.has(vals.recipe);
```

评论区精华

唯一实质性评论来自 gemini-code-assist[bot]：在示例输出中发现 </think> 后缺空格、every100 应为 every 100、Step2/Step3/Step4 缺空格。无后续回复，但 PR 已被 wisclmy0611 批准，因此这些排版问题被视为非阻塞。

风险与影响

风险：

- JSX 中新增的约束逻辑（禁用 / 回退）可能存在遗漏的组合，导致 UI 行为不符合预期。
- 新增的 H100 FP8 基准数据可能在未来版本中过时，需要定期同步。
- 当前无 JSX 的单元测试覆盖。

影响：

- 用户获得更清晰的部署配置界面和完整的 H100 FP8 支持。

- 团队维护成本略有增加，但结构更可扩展。
- 无运行时影响。

关联脉络

此 PR 是 DeepSeek-V4 文档持续演进的一部分，与之前 #25661（添加 FLUX.2-klein-base）、#25988（diffusion 预热）等文档 / 功能 PR 共同完善模型支持矩阵。无直接功能依赖。