# PR #34099 完整报告

sgl-project/sglang

docs: clarify K3 VLM feature transport

合并时间：2026-08-08 19:23

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/34099

## 执行摘要

本 PR 是 K3 VLM 部署文档的澄清性改动：命令选择器新增 VLM Transport 维度，按拓扑动态展示 Auto 的解析结果，并新增显式 CPU 选项；cookbook 正文删除易被误读的静态 B300 推荐命令，改用表格区分 processor-to-scheduler 特征传输与 EPD、PD 传输路径。纯文档与展示层变更，不影响运行时行为。

## 功能与动机

PR body 指出，此前 B300 示例可能被误读为“普适最优”，且模糊了 processor-to-scheduler 特征传输与 EPD、PD 传输路径。作者的目标是让命令选择器显示 Auto 按拓扑如何解析（PD 走 CPU、单节点 B300 走 CUDA IPC、GB200/GB300 有 IMEX 时走 CUDA VMM），并提供一个显式 CPU 选项用于保留 HBM 余量。

## 实现拆解

1. docs/src/snippets/configs/moonshotai/kimi-k3.jsx：在 overlayDims 新增 mmTransport 维度。Auto 选项的 hints 回调按 pdMode 与 hw 分支返回说明文案；CPU 选项通过 flags 注入 --mm-feature-transport cpu。showWhen 限定 pdMode !== 'decode'，因为 decode 节点不处理视觉特征。
2. docs/cookbook/autoregressive/Moonshotai/Kimi-K3.mdx：删除 Recommended high-speed VLM 静态命令与逐条说明，替换为 VLM feature transport 小节与四行表格；VLM compatibility 表格的 MM feature transport 行改为 Processor-to-scheduler features only，明确 EPD 与 PD 各自独立。
3. 精简 Low-HBM VLM 示例：移除冗余的 SGLANG_VIT_ENABLE_CUDA_GRAPH=0 与默认 worker 参数，保留 --mm-feature-transport cpu。
4. 验证：无新增测试；作者在 PR body 声明执行了 node docs/scripts/check_cookbook_configs.mjs、git diff --check 与 pre-commit 钩子。

### docs/src/snippets/configs/moonshotai/kimi-k3.jsx

命令选择器核心改动：新增 mmTransport 覆盖维度，按拓扑返回 Auto 解析 hints，并提供 CPU 选项注入 --mm-feature-transport cpu。这是本 PR 信息展示的核心载体。

```
// overlayDims 中新增的 VLM Transport 维度：让命令选择器显式展示
// `--mm-feature-transport` 在不同拓扑下的解析结果，避免读者误以为
// B300 示例是普适最优配置。
{
```

```
    id: 'mmTransport',
    title: 'VLM Transport',
    default: 'auto',
    // decode 节点不处理视觉特征，因此该维度仅对 Unified / Prefill 显示
    showWhen: (s) => s.pdMode !== 'decode',
    options: [
      {
        id: 'auto',
        label: 'Auto (topology-aware)',
        // hints 按拓扑返回解析说明：目的是澄清 processor 到 scheduler
        // 的特征传输路径，并把它与 EPD / PD 的 KV/KDA 传输区分开
        hints: (s) => {
          if (s.pdMode !== 'unified') {
            return [
              'VLM transport: Auto -> CPU for PD; KV/KDA transfer is separate.',
            ];
          }
          if (s.hw === 'b300') {
            return [
              'VLM transport: Auto -> CUDA IPC (up to 1 GiB HBM; CPU fallback when full).',
            ];
          }
          if (['gb200', 'gb300'].includes(s.hw)) {
            return [
              'VLM transport: Auto -> CUDA VMM with IMEX, otherwise CPU (up to 1 GiB HBM).',
            ];
          }
          return ['VLM transport: Auto -> CPU on this topology.'];
        },
      },
      {
        id: 'cpu',
        label: 'CPU (save HBM)',
        // 显式 CPU 选项：保留 HBM 余量，等价于 --mm-feature-transport cpu
        flags: ['--mm-feature-transport cpu'],
        hints: ['VLM transport: CPU; no GPU feature pool.'],
      },
    ],
  },
```

## 评论区精华

该 PR 无 review 评论与讨论线程，所有决策均来自 PR body 的动机说明，即消除 B300 示例的歧义并区分三条传输路径。

## 风险与影响

- 文档与实现一致性：mmTransport 的 Auto 分支（CUDA IPC、CUDA VMM、CPU）必须与实现侧自动选择逻辑保持同步，否则文档会误导部署。当前实现逻辑来自 #33936。

- 命令生成回归：新增 overlay dimension 可能改变命令选择器其他维度的组合渲染，jsx 无单元测试覆盖，回归风险主要在展示层。
- 用户习惯：移除静态推荐命令后，部分用户可能找不到快速启动命令，但表格与 picker 已提供等价信息。
- 影响范围：仅影响 K3 cookbook 文档与命令选择器渲染，不影响运行时行为。

## 关联脉络

本 PR 是 #33936“feat(vlm): auto-select CUDA VMM on multi-node MNNVL”的文档配套。该实现为多节点 MNNVL 自动选择 CUDA VMM，并对 K3 等多模态模型生效；本 PR 将 Auto 的拓扑感知行为写进命令选择器与 cookbook。后续若传输策略继续演进（如新增拓扑或 IMEX 行为变化），需要同步更新此处 hints 分支。