

PR #39373 完整报告

sgl-project/sglang

[Diffusion] docs: give the RTX 5090 its own H3 recipe, measured on a physical desktop

合并时间: 2026-09-15 09:15

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/39373>

执行摘要

该 PR 是纯文档变更，但核心价值有两层：一是为 RTX 5090 单卡在 MiniMax-H3 部署矩阵中新增独立的 32 GB 消费级档位，修正了此前 5090 因不匹配任何档位而错误继承双卡 TP2 配方的问题；二是带来一份全部硬件轴均为真实物理桌面（无 allocator 上限、无 psutil 补丁、无页缓存保底）的实测数据，并据此把磁盘吞吐提示从经验值改写为可计算的定量模型。改动不涉及任何运行时代码，影响范围是 H3 文档的部署命令渲染与配方提示。

功能与动机

PR body 把动机讲得很清楚：RTX 5090 单卡落不进 CONSUMER_12G/16G/24G、工作站档位与 UNIFIED_128G 中的任何一个，CONSUMER_SINGLE 永远匹配不到它，于是默认配方解析到为双卡实测的通用 offload 旗标——"二十层驻留 DiT 在 TP2 下每 GPU 只有半层，放到单卡上是 32 GB 板卡上的 24.6 GiB 权重"；同时缺少 `video_vae=36`，decoder 每次 decode 都流式读取 36 个块，实测 40.2 s 对比驻留 6.4 s。又因为 5090 不在 CONSUMER_SINGLE 中，Host RAM 选择器被隐藏、提示完全缺失——而这恰是主机内存决定成败的一张卡。

实现拆解

1. 新增 32 GB 档位定义：docs/src/snippets/configs/MiniMaxAI/minimax-h3.jsx 中新增 `CONSUMER_32G = ["rtx5090"]`，并入 `CONSUMER_SINGLE` 与 `CONSUMER_VRAM_16_PLUS`。前者让 5090 单卡首次进入消费级单卡分支，获得 Host RAM 选择器、consumer hints 与单卡配方；后者让 5090 在 ram96 大主机下也获得 4 层 resident 规则与对应提示。
2. 新增 `consumerSingleCard()` 单卡判定：这是本 PR 最关键的逻辑。5090 是唯一有已验证双卡配方的消费级卡（TP2、每 GPU 20 层 resident），双卡选择必须继续走通用 offload 路径；新判定函数用 `Number(s.gpus_per_node) > 1` 排除双卡，保证已有双卡命令字节级不变。auto 下拉框的 flags 与 hints 入口也都从 `CONSUMER_SINGLE.includes` 切换为 `consumerSingleCard`。
3. `consumerFlags()` 增加 32G 分支：当 5090 单卡且主机非 ram96 时，追加 `--dit-layerwise-resident-layers 14`（约 17 GiB），依据是 14 层给 decode 留出空间并缩小主机需要 pin 的流式集合；ram96 大主机仍走已有 4 层规则。
4. `consumerHints()` 提示体系重写：新增 5090 实测提示（60 GB 主机 112.2/112.1 s per request, denoise 5.14-5.17 s/step, decode 6.4 s，对比 ComfyUI 140.9-145.9 s）；把原 "38 s/step 是磁盘瓶颈" 的经验提示替换为可计算模型 `bytes per step = (50 -`

resident DiT layers - pin budget 覆盖层数) x 1.23 GiB; 新增 4 条通用提示: resident层收益的 pin 预算边界条件、按目标形状 warmup、95% pin 预算导致的桌面 swap 压力、旧 38 s/step 结论的归因。

5. cookbook 正文: MiniMax-H3.mdx 新增 "RTX 5090 single card, physical desktop" 章节, 给出完整测量表与三个封顶模拟测不出的结论 (compute wall 而非磁盘、主机内存决定流量规模、两引擎默认命令在这台机器上都无法直接运行); 原 5090 章节改名 "(two cards)" 并增加交叉引用; 修正 ComfyUI --fast-disk 的表述。
6. 验证配套: 无运行时测试 (documentation only); PR body 声称对全部 270 种 "硬件 x GPU 数 x 主机内存" 组合做了渲染前后对比, 仅 5090 单卡命令行变化, 双卡命令字节级不变; minimax-h3.jsx 作为 ESM 可解析、config 可导入求值。

docs/src/snippets/configs/MiniMaxAI/minimax-h3.jsx

H3 部署矩阵的渲染核心, 新增 CONSUMER_32G 档位与 consumerSingleCard() 单卡判定, 重写 consumerFlags 与 consumerHints, 是本 PR 的配置逻辑主体。

```
// 档位定义: 每个 VRAM 档位对应一份配方。
// RTX 5090 是唯一在真实物理桌面 (60 GB 主机 + PCIe 5.0 NVMe) 上
// 实测过的档位, 其余档位的数据来自硬性分配器上限的模拟。
const CONSUMER_12G = ["rtx4070", "rtx5070", "rtx3060"];
const CONSUMER_16G = ["rtx4080", "rtx5080", "rtx5070ti", "rtx4060ti"];
const CONSUMER_24G = ["rtx4090", "rtx3090"];
const CONSUMER_32G = ["rtx5090"];

// 单卡判定: 5090 是唯一有已验证双卡配方的消费级卡 (TP2、每 GPU 20 层驻留),
// 双卡选择必须留在通用 offload 路径上, 只有单卡才走下方档位配方。
function consumerSingleCard(s) {
  return (
    CONSUMER_SINGLE.includes(s.hw)
    && !(CONSUMER_32G.includes(s.hw) && Number(s.gpus_per_node) > 1)
  );
}

function consumerFlags(s) {
  if (UNIFIED_128G.includes(s.hw)) return unified128Flags();
  if (WORKSTATION_96G.includes(s.hw)) return workstation96Flags();
  const flags = [
    "--performance-mode memory",
    "--layerwise-offload-components dit,text_encoder,vae",
    "--layerwise-resident-layers video_vae=36",
  ];
  // 单卡 5090 驻留 14 层 DiT (约 17 GiB): 给 decode 留出空间,
  // 同时缩小主机侧需要 pin 的流式集合, 缓解桌面 swap 压力。
  // ram96 大主机仍走上方 4 层规则, 这里只覆盖 ram32/ram64。
  if (CONSUMER_32G.includes(s.hw) && s.host_ram !== "ram96") {
    flags.push("--dit-layerwise-resident-layers 14");
  }
  return flags;
}
```

```
}
```

评论区精华

该 PR 没有实质性 review 讨论：仅有的两条 comment 分别是 Mintlify preview bot 的预览部署通知与作者自己的 `/tag-and-rerun-ci` 命令，`review_comments` 为 0。所有设计权衡（14 层 resident、ram96 走 4 层、双卡命令不变、磁盘吞吐公式）均由作者在 PR body 中以实测数据与推导形式给出，未引发 reviewer 质疑或返工。

风险与影响

风险：

- 配置矩阵逻辑的连带影响：CONSUMER_SINGLE 与 CONSUMER_VRAM_16_PLUS 同时纳入 32G，会改变 bigHost 分支的提示选择与 ram96 的 resident 规则；auto 下拉框入口函数被替换为 consumerSingleCard，未来新增硬件时若遗漏双卡排除分支，可能再次配方错配。验证靠作者离线渲染对比，不在 CI 中。
- 实测数据的外推边界：cookbook 中 4090 在 32 GB 主机上的性能是推导值而非重新测量，作者明确标注 "that machine has not been re-measured"；磁盘模型预测 49 GiB/request，实测 51.3 GiB，约 5% 误差。
- 文档误导面：hint 文案大量重写，"resident 层只在流式集合超过 pin 预算时减少磁盘流量" 这类带边界条件的结论若被断章取义，可能造成无效配置。
- 无运行时回归面：不涉及生产代码；CI 的 PR Test (Extra) 失败与文档变更无关。

影响：RTX 5090 单卡用户从错误配方（24.6 GiB 驻留 + decode 全程流式约 40 s）切换到正确配方（14 层驻留 + video_vae=36 + Host RAM 选择器 + 磁盘提示），实测 112.2 s/request，较 ComfyUI 快约 30%。对团队确立了 "物理桌面实测 + 可计算模型" 的文档标准，并纠正了被旧路径误导的历史结论。对系统零运行时影响。

关联脉络

PR body 明确引用两个前置修复：#39021（pinned-store 池按精确大小分配，消除 2 的幂取整导致的 pin 预算超卖）与 #39022（courier 以 O_DIRECT 读取流式权重），本 PR 中 "38 s/step 是旧路径问题而非磁盘瓶颈" 的归因正建立在两者之上；#35967 是 decoder fp16 持有 bit-identical 的依据。在同仓库近期 PR 中，#39396、#39406 同属 "部署矩阵 / 配方文档" 演进线，与本 PR 一起把 SGLang 的 diffusion 部署文档推向 "每个硬件组合都有可验证配方" 的目标。