PR #35816 完整报告

sgl-project/sglang

[diffusion] docs: add tuning guide for h3 on consumer-level gpu

合并时间：2026-08-22 23:48

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/35816

MiniMax-H3 消费级 GPU 调优指南（PR #35816）

执行摘要

该 PR 为 MiniMax-H3 补齐消费级 GPU 的部署调优指南，核心论断是：消费级硬件上主机内存才是首要约束，显卡是第二约束。108 GB 的权重（DiT 61.73 GB + text encoder 46.18 GB）决定了没有消费级配置能全部驻留，短 fall 落点直接决定吞吐。配套将文档站 deployment builder 的 H3 配置重构为预算感知的 IIFE 工厂（新增 Host RAM 维度与 14 款消费级 / 工作站显卡），并建立 verified/derived recipe 文档契约与双向校验。全部数字来自 RTX 4090 实测，与 ComfyUI 在同等硬性 12 GiB allocator cap 下对比，输出 bit-identical。

功能与动机

PR body 明确说明：现有 recipe（24 GB 单卡、双 32 GB 卡）都假设主机能容纳全部权重，「No consumer configuration can, and where the shortfall lands decides the throughput」。「On consumer hardware the host is the binding constraint, not the card. The H3 page did not say so.」文档缺失的核心信息有两层：

1. 32 GB 主机内权重无法 pin，每一步需从 checkpoint mapping 同步复制约 60 GiB，这是步进时间的主要来源，NVMe 因此是硬性要求；
2. 用户需要一个「预算感知」的命令生成器，按显卡档位 × 主机内存组合给出可执行的 flags 与实测预期，而不是一份只针对数据中心卡的静态命令。

实现拆解

1. cookbook 新增「Consumer GPU tuning」章节（MiniMax-H3.mdx，+181 行）

- 给出 Recipe A（12 GB VRAM + 32 GB 主机）与 Recipe B（主机内存充裕的快速路径）两条命令；Recipe B 以 4 个驻留 DiT 层换取 6.01 s/step，但需要 16 GB VRAM 与约 112 GB pinned 主机内存。
- 测量方法学两则：主机内存按匿名页 RssAnon 计数而非 VmRSS；VRAM 用 torch.cuda.set_per_process_memory_fraction 硬性 cap 而非 nvidia-smi。
- 启动日志三行自查（host memory available、leaving N GiB of weights on the checkpoint mapping、video_vae 的 host mmap vs host pageable），把机器是否匹配预算的确认提前到第一分钟。
- ComfyUI 对比升级为 fair-cap：两边同权重、同硬 cap 12 GiB，Recipe A 在 32/48/64 GB 主机全请求胜出（235/218/180 s vs 276-302/246-267/194-195 s），并指出

ComfyUI 的 --reserve-vram 是软限制（实测峰值 13.5 GiB）。

2. builder 配置重构为预算感知工厂（minimax-h3.jsx，+185/-23）

- 关键修复：原静态对象引用的顶层常量不被 snippet 系统跨文件携带，曾导致整个 H3 页面空白；改为 IIFE 闭包后常量与 config 同作用域。
- 新增 host_ram overlay 维度（32/48-64/96 GB+）与 14 款消费级 / 工作站显卡，按 12/16/24/48/96 GB VRAM 分层共享测得数据。
- consumerFlags(s) 按「卡档位 × 主机预算」组合生成 flags：12 GB 基础为 video_vae=36 解码期驻留 + 组件 offload；24 GB + 32 GB 主机追加 10 个 DiT 驻留层（实测 10.4 vs 11.6 s/step）；96 GB+ 主机追加 4 个；48 GB 工作站追加 40 个；96 GB 工作站整套 DiT 驻留。
- consumerHints(s) 为每个组合输出实测预期与硬性前提，并提示「启动日志应输出 leaving ... GiB of weights on the checkpoint mapping，否则主机不是你以为的约束」。
- 同步清理了 datacenter recipe 中不适用于消费级的 --dit-offload-prefetch-size 1 与 --enable-torch-compile false。

3. 渲染器适配（_deployment.jsx，+13/-3）

- handleSelect 的 hw 切换分支在继承资源形状之外追加继承新卡的 placement 与 encoder，避免从 RTX 5090 切到 RTX 4090 时保留 resident 选择、生成单卡无法运行的命令。
- recipe 徽标与按钮区分 Verified/Derived 两态，与 unverified 标志联动。

4. 文档契约校验（check_cookbook_configs.mjs，+10/-1）

- 双向断言：普通 verifiedRecipes 条目必须解析为 verified；声明 unverified 的条目若解析为 verified 则直接 fail，防止契约静默漂移。

5. 测试与依赖配套

- 纯文档 / 站点配置 PR，无运行时代码与测试文件；性能数据全部来自单张 RTX 4090 实测（864×480 / 124 帧 / 20 NFE / bf16 / seed 1101）。
- 落地依赖 #35967（fp16-held VAE decoder）合并，PR body 明确注明该依赖已满足后才最终修订 video_vae=36 在 12 GB 上的结论。

docs/src/snippets/_deployment.jsx

部署渲染器适配：硬件切换时继承新卡的 placement 与 encoder，避免生成单卡无法运行的 resident 命令；新增 Derived/Verified recipe 双态徽标与按钮文案，支撑 unverified 契约。

```
// 硬件切换时，builder 不仅需要跟随新卡调整资源形状（节点数、TP 等），
// 还必须重置 placement 与 encoder：它们是按硬件验证的 recipe 事实。
// 若沿用上一张卡的选项，可能生成新卡跑不了的命令——例如在单张
// 消费级卡上 resident 61.7 GB DiT——并错误显示为 unverified。
if (dim === "hw") {
  const currentRecipe = recommendedBuilderRecipe(prev.hw);
  const nextRecipe = recommendedBuilderRecipe(value);
  const resourcesFollowPlatformDefault = !!currentRecipe
    && Number(prev.nodes) === Number(currentRecipe.nodes)
    && Number(prev.gpus_per_node) === Number(currentRecipe.gpus_per_node);
```

```
    next = {
      ...next,
      nodes: resourcesFollowPlatformDefault
        ? (nextRecipe?.nodes ?? next.nodes)
        : next.nodes,
      gpus_per_node: resourcesFollowPlatformDefault
        ? (nextRecipe?.gpus_per_node ?? next.gpus_per_node)
        : next.gpus_per_node,
      topology_mode: "auto",
      tp_size: resourcesFollowPlatformDefault
        ? (nextRecipe?.tp_size ?? 1)
        : next.tp_size,
      placement: resourcesFollowPlatformDefault
        ? (nextRecipe?.placement || "auto")
        : next.placement,
      encoder: resourcesFollowPlatformDefault
        ? (nextRecipe?.encoder || "auto")
        : next.encoder,
    };
  }
```

docs/scripts/check_cookbook_configs.mjs

为 derived/unverified recipe 建立双向校验契约：普通条目必须解析为 verified，unverified 条目不得解析为 verified，防止文档契约静默漂移，是本次文档工程化的关键配套。

```
  // recipe 可以携带 unverified: true：它只提供该卡的默认资源形态，
  // 不声称做过验证运行，且必须按此解析。这里做双向断言：
  // 普通条目必须解析为 verified；声明 unverified 的条目一旦解析为
  // verified 就立即 fail，防止文档契约静默漂移。
  if (recipe.unverified) {
    const resolved = validateResolved(selection, `verifiedRecipes[${index}]`);
    if (resolved && resolved.builder.verification?.serve === "verified") {
      fail(where, `verifiedRecipes[${index}] is declared unverified but resolves as verified`);
    }
  } else {
    validateResolved(selection, `verifiedRecipes[${index}]`, true);
  }
```

评论区精华

该 PR 没有任何 GitHub review 评论（comments_count = 0，review_comments_count = 0），实质讨论发生在 16 个 commit 的演进中，且多次是作者对自身错误结论的公开修正：

> Recipe B 的 VRAM 门槛修正（ed40a20）：最初建议 12 GB 可用，实测 4 个驻留 DiT 层在 12 GiB cap 下 OOM，必须 16 GB。
>
> ComfyUI 对比结论反转（3f4bfe14）：早期声称「ComfyUI 不能在 12+32 下跑 bf16」其实是测量错误——ComfyUI 内存管理器读取系统 RAM 自适应，用 psutil patch 模拟 32 GiB 世界后它反而以 13.0 s/it 领先当时的 Recipe A（17-18 s/it）。

> 全请求胜出依赖 decode 优化（187c54c7）：denoise 领先但流式 decode 需 209 s，「三分之二解码器为 decode-only 驻留、结束后释放」才在 12 GiB cap 内把整请求压到 279-283 s。
>
> 页面空白事故（34ba7446）：配置常量放模块顶层导致 snippet 渲染整页空白——docs 站点构建系统的真实陷阱，IIFE 修复。
>
> 数字边界诚实声明：正文写明 12% 的请求时间波动跟踪 host load、更小差异不作数；32 GB 图形是在 2 TB 主机上取到的乐观值，真实 32 GB 机器会因 re-read 更慢。

风险与影响

风险

- 数字时效性：所有数字是特定硬件组合（RTX 4090 + 2 TB 主机）的实测，随 layerwise offload 与解码器相关运行时优化（#35967、#35734 等）持续演进可能过时，需要专人维护同步。
- builder 配置回归风险：minimax-h3.jsx 是 docs 站点运行时渲染的代码，曾发生整页空白事故；本次重构未增加自动化测试，仅靠 check_cookbook_configs.mjs 兜底，而该脚本不校验 flags 合法性。
- unverified 契约误报风险：新增双向断言会让「声明与解析不一致」直接 CI fail，虽有意的护栏，但要求作者正确理解 verify 解析语义。
- 用户误导风险：12 款新硬件条目共享档位数据，30 系卡步进时间高于 40 系实测、48/96 GB 工作站卡为 derived 未验证，用户若不细读 hints 可能误判自身硬件表现。

影响

- 用户面：消费级 H3 用户获得首份可决策的部署指南——预算匹配的 flags、实测预期、ComfyUI 公平对比与启动日志自查。
- 站点面：deployment builder 新增 Host RAM 维度与 14 款显卡，H3 页面交互能力显著扩展，并修复了整页空白 bug。
- 团队 / 流程：建立 verified/derived recipe 文档契约与双向校验脚本，为后续新增硬件提供可扩展模板。
- 运行时影响为零，不涉及 python/sglang 任何代码。

关联脉络

该 PR 是「消费级硬件跑 H3」这条功能线上的文档收口，与多个运行时优化 PR 配对演进：

- 前置依赖 #35967（fp16-held VAE decoder）：video_vae=36 在 12 GB 上成立的前提，PR 等它合并后才落地；
- 36051（guard anonymous-host budget）与本 PR 的 RssAnon 方法学同源，同属「主机内存是首要约束」这一主题；
- 35734（layerwise 组件非层权重暂存）与 consumerFlags 的驻留层策略互为依据；
- 36034（清理启动 /offload 日志）为文档「Reading the startup log」一节提供日志行；
- 36032（5090 consumer case CI）与本 PR 的消费级实测数字互相印证，构成「文档 + 运行时 + CI」三位一体的消费级 H3 支持体系。