

PR #39244 完整报告

sgl-project/sglang

docs: refresh the VDN-H3 on B200 numbers in the MiniMax-H3 cookbook

合并时间: 2026-09-14 11:50

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/39244>

执行摘要

- 一句话: 重测 VDN-H3 B200 基准并更新 overlay
- 推荐动作: 值得快速浏览。虽是文档 PR, 但包含两类可复用信息: VDN-H3 在 B200 上的性能标尺与完整部署配方 (对硬件选型与容量规划有决策价值); overlay pin 暴露出 "模型仓库文件名变更会破坏既有部署" 的问题, 说明 overlay 机制缺少自动化校验。对需要发布 benchmark 数据的团队, 本 PR 的测量方法 (布局扫描、声明波动范围、区分并行效率与内核性能、未重测即标注) 值得借鉴。

功能与动机

PR body 明确指出原表格已过期: "The **VDN-H3 on B200** table in the MiniMax-H3 cookbook is out of date: the SGLang rows predate the current **mxfp8** mapping and the current memory policy, and the OpenVDN reference rows were taken at one assumed Ulysses layout." 此外, 提交列表第二条给出了功能性动机: OpenVDN 将 `adapter_config.json` 重命名为 `adapter_spec.json`, 旧的 pinned overlay 仍要求旧文件名, 导致全新 **sglang serve --model-path OpenVDN/vdn-minimax-h3** 在加载任何权重前即中止, 需要同步提升 overlay revision 才能恢复部署。

实现拆解

1. 更新 overlay 版本钉 (唯一源码变更): `python/sglang/multimodal_gen/runtime/utils/model_overlay.py` 的 `BUILTIN_MODEL_OVERLAY_REGISTRY` 中 `OpenVDN/vdn-minimax-h3` 条目 `overlay_revision` 由 `0ad315a05b914c4003af4d26152d288c2506a609` 提升至 `7de18275dddfe59da36a234e222bcdd274963bc3`。新 overlay 读取 OpenVDN 重命名后的 `adapter_spec.json` (并回退兼容旧名), `manifest` 强制要求该版本; 提交说明在 `8xB200` 上验证融合输出与旧 overlay 字节一致。注意 overlay 指纹随之变化, 已缓存部署会触发一次权重重新物化。
2. 重写 VDN-H3 章节引言: `docs/cookbook/diffusion/MiniMax/MiniMax-H3.mdx` 改用 OpenVDN 官方术语描述 Video DeltaNet 混合注意力架构 (滑动窗口 softmax 分支 + Video Delta Attention 线性分支), 删除 FastH3 交叉引用与内部形状细节, 明确 8-NFE DMD2 distill、4.3 GB 线性分支与 FL2VA 分区。
3. 更新启动配方与性能表: 默认配方改为 `8xB200 (--num-gpus 8)`, 并说明 `4/2/1` 对应较小 Ulysses 配方; 重测后 `8xUlysses8` 在线 mxfp8 从 0.93 降至 0.88 s/NFE, `1x` 从 7.49

降至 6.03, 4x 由 per-channel fp8 的 2.34 降至在线 mxfp8 的 1.67; 新增 2xUlysses2 行 (3.30 s/NFE), 使系列完整为 1/2/4/8; RTX PRO 6000 段落引用的 B200系列同步改为 6.03 / 3.30 / 1.67 / 0.88。

4. OpenVDN 参考行重测与布局扫描: 不再假设单一 Ulysses 分割, 而是在每个 GPU 数下扫描 parallel.softmax_ranks 布局 (8x 5+3 为 1.40、4x 3+1 为 2.59、2x 1+1 为 5.04 等), 并精确复现其公开发布的 8x 数字 (1.40 s/NFE), 确认该行保留。
5. 删除失效内容并重写对比段落: 单 GPU 下 DiT 在自动内存策略下已可常驻 (128 GB 量化 / 156 GB BF16, 卡为 183 GB), 两条 --layerwise-offload-components 行及 "BF16 DiT 在该片段长度下无法常驻" 的说法被删除; 对比段落改为: 等卡数下 SGLang 快 1.48x/1.55x/1.60x (2/4/8 卡), 单卡两栈差距在 4% 内, 差异本质是并行扩展效率 (86-91% vs 56-64%) 而非单 GPU 内核性能; 原 "单卡差距来自在线 FP8 GEMM 路径" 的说法被删除, 因为单卡两种精度下 SGLang 均已更快。未重测的 per-DiT-block 注意力拆分 (117 ms) 明确标注为未重测。

关键文件:

- docs/cookbook/diffusion/MiniMax/MiniMax-H3.mdx (模块 文档; 类别 docs; 类型 documentation): 核心变更文件, 重写 VDN-H3 章节: 重测后的 B200 全表、新增 2x Ulysses2 行、OpenVDN 布局扫描、删除失效内存段落、重写对比结论。
- python/sglang/multimodal_gen/runtime/utils/model_overlay.py (模块 覆盖层; 类别 source; 类型 data-contract; 符号 BUILTIN_MODEL_OVERLAY_REGISTRY): 唯一的源码变更: 将 OpenVDN/vdn-minimax-h3 的 overlay_revision 升级以读取重命名后的 adapter_spec.json, 修复全新部署启动失败, 是本次变更中唯一有运行时影响的地方。

关键符号: 未识别

关键源码片段

python/sglang/multimodal_gen/runtime/utils/model_overlay.py

唯一的源码变更: 将 OpenVDN/vdn-minimax-h3 的 overlay_revision 升级以读取重命名后的 adapter_spec.json, 修复全新部署启动失败, 是本次变更中唯一有运行时影响的地方。

```
# 内置扩散模型 overlay 注册表: 把 Hugging Face 模型映射到 overlay 仓库
# 及其锁定 revision, 运行时据此物化模型结构与权重布局。
BUILTIN_MODEL_OVERLAY_REGISTRY: dict[str, dict[str, Any]] = {
    # OpenVDN 将 VDN-H3 的 adapter_config.json 重命名为 adapter_spec.json,
    # 旧 pinned overlay 仍要求旧文件名, 导致新部署在加载权重前即中止。
    # kevin-mi/VDN-H3-overlay@7de18275 读取新名并回退旧名, 已在 8xB200
    # 上验证融合输出与旧 overlay 字节一致。
    "OpenVDN/vdn-minimax-h3": {
        "overlay_repo_id": "kevin-mi/VDN-H3-overlay",
        "overlay_revision": "7de18275dddf59da36a234e222bcdd274963bc3",
    },
}
```

评论区精华

本 PR 的 comments 与 review comments 均为 0，没有评讨论，关键决策沉淀在提交信息中：

- kevin-mii 在提交 2 中说明 overlay 升级的必要性：OpenVDN 重命名 adapter_config.json 为 adapter_spec.json 后，旧 pinned overlay 导致全新部署启动即中止；新 overlay 7de18275 兼容新旧命名，且 " 融合输出与旧 overlay 字节一致 "，保证基准数据可复现。
- 提交 3、4、5 (kevin-mii, Claude Fable 5.1 参与) 逐步收紧文档：删除未发布方案 (block-sparse kernel、tried-and-not-shipped 布局) 与未重测的 per-block profile，避免读者误读为当前状态；将 Attention Backends 链接指向 --attention-backend-config 键列表。
- 方法论取舍：重复臂跨节点波动约 3%，PR body 明确声明表格数字来自单次 campaign，正文已写入该不确定性。
- 暂无高价值评论线程

风险与影响

- 风险：
 1. overlay 指纹变化触发权重重新物化：_compute_overlay_fingerprint 依据 overlay 目录内容计算指纹，revision 变更使指纹改变，已有 VDN-H3 部署会重新下载并融合权重 (约 62 GB 写入)，需在 SGLANG_DIFFUSION_CACHE_ROOT 预留至少 90 GB 空间 (文档已注明)。
 2. 源码变更无自动化测试：PR body 声称 docs only，但实际包含 model_overlay.py 的 1 行数据契约变更；虽有 8x200 手工验证与字节一致性结论，仓库内没有对应自动化测试覆盖该 overlay 条目。
 3. 数据时效性：所有性能数字来自单次 campaign，同一节点上重复臂波动约 3%；117 ms 注意力拆分未重测，文档虽已标注但用户可能误用。
 4. 文档与软件版本强耦合：表格依赖特定版本行为 (在线 MXFP8 默认、auto memory policy、Ulysses 布局)，后续代码变更可能再次使文档过期，这正是本 PR 解决的原始问题。- 影响：对用户：部署 VDN-H3 / MiniMax-H3 的用户获得最新、正确的 B200 基准与启动配方，且修复了全新部署因 overlay 文件名不匹配而在加载权重前中止的故障；B200 系列数据同时被 RTX PRO 6000 段落引用，硬件选型参考随之更新。对系统：除 overlay revision 外无运行时行为变化，已有部署会因指纹变化触发一次性权重重新物化。对团队：5 次提交中 4 次由 kevin-mii 与 Claude 协作完成，展示了大文档重写的迭代方式；PR Test 通道标红但未阻塞合并。- 风险标记：overlay 版本变更触发权重重新物化，性能数据来自单次测量，源码变更无自动化测试，文档与软件版本强耦合

关联脉络

- PR #39292 [Diffusion] Don't route an unreadable checkpoint into the native fallback: 同属 diffusion 模型加载健壮性主题：该 PR 阻断不可读 checkpoint 进入 fallback，本 PR 修复 overlay 因 adapter 文件名变更导致加载失败，均围绕加载路径的失败模式。
- PR #39291 [Diffusion] Make the SP sequence gather pass contiguous shards: 同属 diffusion 运行时功能线，涉及 multimodal_gen 运行时层与注意力后端行为，与本 PR 的模

型加载与 B200 部署配方同属 diffusion 支持面。