

PR #36028 完整报告

sgl-project/sglang

[docs] Add MiniMax H3 checkpoint format table

合并时间: 2026-08-26 09:18

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36028>

执行摘要

- 一句话: 新增 MiniMax H3 checkpoint 格式总表, 整合量化 CLI 文档
- 推荐动作: 值得查阅但不必读代码: 使用 MiniMax-H3 或 SGLang-Diffusion 量化时, 这份表格是最新的格式 -CLI 对照; 文档维护者可借鉴“来源驱动表格 + 单一事实源 + 锚点复用”的结构。合并前建议回看 Extra CI 失败, 并抽查删除行是否覆盖了 GGUF 等细节。

功能与动机

PR body 明确目标: consolidate supported MiniMax H3 checkpoint and adapter formats into one compact table、show the canonical CLI addition for each format、distinguish native model execution from external kernels and load-time conversion。此前模型级 checkpoint 资料分散在多个第三方来源, 量化文档按参数罗列路径, 用户容易按精度名而不是按手中资源选 CLI; quantization.mdx 新表述直接写明 Choose the option from the source you have, not from the precision name。

实现拆解

1. 在 cookbook 中新增 checkpoint 总表并修正任务描述:
docs/cookbook/diffusion/MiniMax/MiniMax-H3.mdx 新增 “Checkpoint and adapter formats” 小节, 用 Scope / Format or variant / Add to the base command / Contract 四列覆盖官方 mixed BF16/FP32、官方 Diffusers 布局、AdaLN-pruned Diffusers、BF16 safetensors、Comfy FP8/MXFP8 等来源, 并给出叠加在 builder 输出命令之上的 CLI 片段; 同时把 ref2va 描述扩展为可结合 first/last keyframes, 且必须至少保留一个 reference condition。该文件净删 228 行, 把散落的细节并入表格。
2. 重构量化指南为来源驱动表格: docs/docs/sglang-diffusion/quantization.mdx 将一长串路径定义改写为 Source or operation / CLI / Result 三列, 明确 --component-paths. 替换 config 与权重、--component-weights-paths. 仅替换权重并保留 base config, 并把 --transformer-path、--transformer-weights-path、--quantization 降级为主 DiT 便捷别名。
3. 更新 CLI API 文档示例并交叉引用: docs/docs/sglang-diffusion/api/cli.mdx 把示例中的具体仓库路径改为 COMPONENT_REPO_OR_DIRECTORY 与 WEIGHTS_FILE_OR_REPO_FILE 占位符, 并新增指向 MiniMax-H3 表格的链接, 声明 model cookbook 是模型级 checkpoint 示例的单一事实源。

4. 兼容矩阵增加导航入口：docs/docs/sglang-diffusion/compatibility_matrix.mdx 在 MiniMax-H3 行追加“Checkpoint formats”锚点链接。
5. 测试与 CI：无源码与测试改动，PR body 说明 documentation-only change; no GPU test requested, 仅跑 changed-file pre-commit; pr-states 显示 Extra 测试失败 (:x:) 但仓库无相关 review, 需合并前确认为无关失败。

关键文件：

- docs/cookbook/diffusion/MiniMax/MiniMax-H3.mdx (模块 模型文档; 类别 docs; 类型 docs-content) : 本 PR 核心: 新增 checkpoint 与 adapter formats 总表, 修正 ref2va 与 fl2va 的关系描述, 并删除 228 行散落说明。
- docs/docs/sglang-diffusion/quantization.mdx (模块 量化文档; 类别 docs; 类型 docs-content) : 将路径式 CLI 列表重构为“来源—CLI—结果”三列表格, 明确 component-paths 与 component-weights-paths 语义边界, 并把 --transformer-path 等降级为主 DiT 便捷别名。
- docs/docs/sglang-diffusion/api/cli.mdx (模块 命令文档; 类别 docs; 类型 docs-entry) : 示例从具体仓库改为通用占位符, 并新增到 MiniMax-H3 表格的引用, 声明 cookbook 是模型级 checkpoint 示例的事实源。
- docs/docs/sglang-diffusion/compatibility_matrix.mdx (模块 兼容矩阵; 类别 docs; 类型 docs-entry) : 兼容矩阵中 MiniMax-H3 行增加“Checkpoint formats”锚点链接, 作为新表格的导航入口。

关键符号：未识别

关键源码片段

docs/cookbook/diffusion/MiniMax/MiniMax-H3.mdx

本 PR 核心：新增 checkpoint 与 adapter formats 总表, 修正 ref2va 与 fl2va 的关系描述, 并删除 228 行散落说明。

```
<!-- 本 PR 核心整理单元：把 Hugging Face 与 ModelScope 上散落的 H3 checkpoint  
来源汇总为一张表，每行给出叠加在 builder 输出命令上的 CLI 片段。  
原表还包含 ConvRot INT8 / W4A8 / W4A4 等更多量化行，此处仅转录  
patch 中完整可见的条目。 -->
```

```
### Checkpoint and adapter formats
```

Start with the command emitted by the builder. Every row below is an overlay on the same native SGLang pipeline; component repositories contribute their own config and weights, while weight files retain the base component config.

```
| Scope | Format or variant | Add to the base command | Contract |  
| --- | --- | --- | --- |  
| Full model | Official mixed BF16/FP32 | `--model-variant fl2va` or `--model-variant ref2va` |  
Lossless reference and consistency GT path. |  
| DiT | Official Diffusers component layout | `--component-paths.transformer MiniMaxAI/  
MiniMax-H3/transformer` (`fl2va`) or `.../transformer_ref` (`ref2va`) | Loads the official
```

```
component through the native SGLang graph; no Diffusers runtime fallback. |
| DiT | [AdaLN-pruned Diffusers component](https://huggingface.co/multimodalart/MiniMax-H3-Pruned) | `--component-paths.transformer multimodalart/MiniMax-H3-Pruned/transformer` or `../transformer_ref` | Approximate curve-AdaLN architecture; its config and basis metadata are loaded natively. |
| DiT | [Full or AdaLN-pruned BF16 safetensors](https://huggingface.co/Comfy-Org/MiniMax-H3/tree/main/diffusion_models) | `--component-weights-paths.transformer OWNER/REPO/path/FILE.safetensors` | Weight-only override. Match the FL2VA/Ref2VA partition; pruned or dtype-converted exports are approximate. |
| DiT | Comfy FP8 or self-describing MXFP8 safetensors | `--component-weights-paths.transformer OWNER/REPO/path/FILE.safetensors` | Per-layer metadata selects static/dynamic FP8 or MXFP8 automatically. |
```

<!-- 同文件另一处语义修正：ref2va 现在允许混入 fl2va 的 first/last keyframes，但仍要求至少一个 reference condition。 -->

```
| Reference to video and audio | `ref2va` | Image, video, and audio references, optionally combined with first/last keyframes |
```

docs/docs/sglang-diffusion/quantization.mdx

将路径式 CLI 列表重构为“来源—CLI—结果”三列表格，明确 component-paths 与 component-weights-paths 语义边界，并把 --transformer-path 等降级为主 DiT 便捷别名。

<!-- 量化文档的关键重构：把原先按参数罗列的路径清单改写为“来源—CLI—结果”三列，纠正“按精度名猜选项”的误导。 -->

```
| Source or operation | CLI | Result |
| --- | --- | --- |
| Complete component repo/directory | `--component-paths.<component> <source>` | Replaces that component's config and weights. |
| Weight file/directory for an existing component | `--component-weights-paths.<component> <source>` | Keeps the base component config and replaces only its weights. |
| Online quantization of unquantized weights | `--component-quantizations.<component> <method>` | Constructs a supported quantized implementation during load; use `--component-quantization-ignored-layers.<component>` for exclusions. |
| Primary DiT convenience spellings | `--transformer-path`, `--transformer-weights-path`, `--quantization` | Aliases for pipelines with one primary DiT; they remain supported but are not generic component names. |
```

<!-- 原表还包含 Causal KV-cache compression 一行 (--kv-cache-quant) , patch 摘要在此截断，未完整转录。 -->

评论区精华

仓库中没有 reviewer 评论，唯一评论是作者 mickqian 自己触发的 /tag-and-rerun-ci，因此没有可提炼的技术交锋。唯一值得注意的是 pr-states 中 Latest PR Test (Extra) 失败 (:x:)，但没有任何讨论记录，推测与纯文档变更无关。

- 暂无高价值评论线程

风险与影响

- 风险：1) 文档与实现一致性：quantization.mdx 把 CLI 语义浓缩成表格，纯文档变更没有自动化校验，若与 server_args 实际行为不一致会误导用户，建议对照参数实现复核。2) 信息丢失：净删 320 行（其中 MiniMax-H3.mdx 删 228 行），需确认 GGUF 支持范围、Hub 子目录写法等细节被表格完整接管。3) 外部链接漂移：表格引用 multimodalart、Comfy-Org、Winnougan、Merserk 等第三方 Hugging Face 仓库，后续改名或下架会造成文档失效。4) CI 噪音：Extra 测试失败无讨论记录，作为文档 PR 大概率无关，仍建议合并前确认。
- 影响：对用户：MiniMax-H3 与量化使用者获得单一权威参考，按来源选 CLI 而非按精度名猜选项，降低误配概率；对系统：纯文档变更，零运行时影响；对团队：cookbook 成为模型级 checkpoint 示例的单一事实源，CLI API 与兼容矩阵通过锚点引用，长期维护面更小。
- 风险标记：纯文档变更，删除信息量大，文档与 CLI 实现一致性需人工核对，外部链接易漂移，Extra CI 失败未确认

关联脉络

- PR #35850 [Diffusion][minimax-h3] Restrict MiniMax-H3 SubBlock sparsity to video queries: 同一 MiniMax-H3 功能线的近期变更，本 PR 是其 checkpoint 选型文档的收口。
- PR #36249 [diffusion] feat: support out-of-tree torch.compile backends: SGLang-Diffusion 运行时组件机制持续演进，本文档中 --component-* 口径与之一脉相承。