

PR #27496 完整报告

sgl-project/sglang

Update SGLang diffusion skills

合并时间: 2026-06-08 12:22

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27496>

执行摘要

- 一句话: 更新扩散技能目录, 移除过时技能, 刷新基准脚本
- 推荐动作: 值得关注 if 您从事 SGLang diffusion 模型开发或基准测试工作。建议阅读 `bench_diffusion_denoise.py` 的模型预设更新和 `existing-fast-paths.md` 的新融合模式, 了解当前支持状态。总体来说, 这是一个维护型 PR, 面向开发者工具, 不涉及核心运行时变更。

功能与动机

根据 PR body, 本 PR 旨在清理过时的 AKO4ALL kernel 技能 (由于不再维护), 并刷新所有扩散相关技能以反映最新的模型支持状态和基准测试配置, 确保新模型和 GPU 测试覆盖被纳入技能文档和脚本中。

实现拆解

1. 删除整个 `sglang-diffusion-ako4all-kernel` 技能目录 (`SKILL.md`、`ensure_ako4all_clean.sh`、`ako-loop.md`), 并清除其他技能中对 AKO4ALL 的引用。
2. 更新 benchmark 脚本 `bench_diffusion_denoise.py`: 扩大 `GATED_MODELS` 包含 `flux2-klein` 系列; 扩展 `MODELS` 字典添加来自注册表和 GPU 测试的当前源码跟踪预设 (`qwen-image`、`qwen-edit-2509`、`flux2-klein`、`cosmos3-nano-t2i` 等); 修复 `env` 变量查找时 `cfg` 未定义的 bug。
3. 更新 `existing-fast-paths.md` 文档, 添加新的融合模式 (`fuse_layernorm_scale_shift_gate`、`varlen USP attention` 等) 和相关测试和基准文件。
4. 更新 `benchmark-and-profile.md` 文档, 明确预设分类 (夜间对齐、当前源码 `extras`、纯技能压力预设)。
5. 微调其他技能的 `SKILL.md` (`performance`、`add-model`、`ModelOpt`、`benchmark-profile`) 以更新手页指南和引用。

关键文件:

- `python/sglang/multimodal_gen/.claude/skills/sglang-diffusion-benchmark-profile/scripts/bench_diffusion_denoise.py` (模块 基准脚本; 类别 `source`; 类型 `core-logic`): 基准测试核心脚本, 修改最大 (+157/-22): 扩展模型预设、更新 `GATED_MODELS`、修复 `env` 变量查找 bug

- python/sglang/multimodal_gen/.claude/skills/sglang-diffusion-benchmark-profile/existing-fast-paths.md (模块 快速路径文档; 类别 docs; 类型 documentation) : 快速路径文档更新, 添加新融合模式和对应测试文件
- python/sglang/multimodal_gen/.claude/skills/sglang-diffusion-ako4all-kernel/scripts/ensure_ako4all_clean.sh (模块 AKO4ALL 脚本; 类别 other; 类型 deletion) : AKO4ALL 环境清理脚本, 被整个删除, 表明关闭 AKO4ALL kernel 优化路径
- python/sglang/multimodal_gen/.claude/skills/sglang-diffusion-ako4all-kernel/SKILL.md (模块 AKO4ALL 技能文档; 类别 docs; 类型 deletion) : AKO4ALL 技能文档, 被删除, 代表整个技能的移除
- python/sglang/multimodal_gen/.claude/skills/sglang-diffusion-benchmark-profile/benchmark-and-profile.md (模块 基准文档; 类别 docs; 类型 documentation) : 基准测试文档更新, 重新分类预设, 并更新手页指南

关键符号: 未识别

关键源码片段

python/sglang/multimodal_gen/.claude/skills/sglang-diffusion-benchmark-profile/scripts/bench_diffusion_denoise.py

基准测试核心脚本, 修改最大 (+157/-22): 扩展模型预设、更新 GATED_MODELS、修复 env 变量查找 bug

```
# bench_diffusion_denoise.py (head version with bugfix applied)

# --- 新增的当前源码跟踪预设 (Current-source extras) ---
# 这些预设来自注册表或 GPU 测试案例, 但尚未纳入夜间比较配置。
"qwen-image": {
    "path": "Qwen/Qwen-Image",
    "prompt": "A futuristic cyberpunk city at night, neon lights reflecting on wet streets",
    "extra_args": ["--width=1024", "--height=1024"],
},
"qwen-edit-2509": {
    "path": "Qwen/Qwen-Image-Edit-2509",
    "prompt": "Make the cat wear a red hat",
    "image_path": str(ASSET_DIR / "cat.png"),
    "extra_args": ["--width=1024", "--height=1024"],
},
"flux2-klein": {
    "path": "black-forest-labs/FLUX.2-klein-4B",
    "prompt": "...",
    "extra_args": ["--width=1024", "--height=1024", "--dit-layerwise-offload", "false"],
},
# ... ( 还有其他 extras, 如 cosmos3-nano-t2i 等 )

# --- run_benchmark_once 中的 env 修复 (根据 review 建议) ---
def run_benchmark_once(model_key, ...):
    env = os.environ.copy()
```

```
env.setdefault("FLASHINFER_DISABLE_VERSION_CHECK", "1")
# 修复：必须先获取模型配置字典，否则 cfg 未定义
cfg = MODELS[model_key] # 新增此行
for key, value in cfg.get("env", {}).items():
    env.setdefault(key, str(value))
```

评论区精华

在 review 中 gemini-code-assist[bot] 指出 bench_diffusion_denoise.py 中的 run_benchmark_once 函数中变量 cfg 未定义，会导致 NameError。作者在第 3 个 commit 'Fix diffusion benchmark env preset lookup' 中修复了此问题。此外，没有其他讨论。

- bench_diffusion_denoise.py 中 cfg 变量未定义导致 NameError (correctness): 作者在第 3 个 commit 中修复了此问题，在代码中添加了 MODELS[model_key] 赋值。

风险与影响

- 风险：风险较低。删除过时技能可能影响依赖它的开发者工作流，但该技能标记为过时且不再维护。benchmark 脚本的改动经过 review 和修复，新增的模型预设需要进行验证（PR body 提供了验证步骤）。主要风险在于新预设是否匹配实际支持状态，但这些预设来自当前注册表和测试，所以风险可控。
- 影响：对用户：删除 AKO4ALL 技能后，开发者需要直接使用 AKO4ALL 仓库而非通过技能引导；新增的模型预设使基准测试更全面。对系统：脚本质量提升，支持更多模型。对团队：维护成本降低（清理过时内容）。整体影响中等偏低。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR