PR #24273 完整报告

sgl-project/sglang

[SKILLS] Tiny upgrade diffusion skills

合并时间：2026-05-02 22:04

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24273

执行摘要

- 一句话：更新 diffusion benchmark 预设并新增 LTX-2.3 支持
- 推荐动作：对于使用 diffusion benchmark 的团队建议尽快同步本 PR，确保基准测试的标准化。同时建议后续技能更新保持这种同步模式，并考虑自动化 weekly watchlist 更新。

功能与动机

保持 bench_diffusion_denoise.py benchmark 脚本与 nightly 对比配置 comparison_configs.json 一致，新增 LTX-2.3 模型 preset，并为开发者提供正在进行的相关 PR 可见性，避免重复工作。

实现拆解

1. 更新 benchmark 脚本：在 bench_diffusion_denoise.py 中新增 ltx23-ti2v-two-stage preset，修改原 ltx2 preset 参数（分辨率从 1536x1024 改为 768x512，启用 CFG parallel，改为 2 GPU 等），添加表格宽度常量并调整打印格式。
2. 同步文档：在 benchmark-and-profile.md 中更新 preset table，对齐 nightly 配置；在 existing-fast-paths.md 中添加 GroupNorm+SiLU 融合、QK norm rope 等记录以及开放 PR 监察清单。
3. 添加 Open PR Watchlist：在 ModelOpt-quant SKILL.md 和 ako4all-kernel SKILL.md 中记录相关开放 PR 编号，指导开发者优先参考已有工作。
4. 更新 Cache-DiT 说明：在 diffusion-performance SKILL.md 中重写 Cache-DiT 使用方法，区分原生 diffusers 后端。
5. 调整表格格式：根据 review 反馈统一表格分隔符长度，使 catalog 和 results 表格对齐。

关键文件：

- python/sglang/multimodal_gen/.claude/skills/sglang-diffusion-benchmark-profile/scripts/bench_diffusion_denoise.py（模块 扩散基准；类别 source；类型 core-logic；符号 print_model_catalog, print_results_table, CATALOG_TABLE_WIDTH, RESULTS_TABLE_WIDTH）：核心 benchmark 脚本，修改了 model presets 和打印格式，新增 LTX-2.3 preset，调整表格宽度，使预设与 nightly 对齐。
- python/sglang/multimodal_gen/.claude/skills/sglang-diffusion-performance/SKILL.md（模块 性能文档；类别 docs；类型 documentation）：更新了 Cache-DiT 说明，区分原生和 diffusers 后端使用方式，并添加 Open PR Watchlist。

- python/sglang/multimodal_gen/.claude/skills/sglang-diffusion-benchmark-profile/existing-fast-paths.md（模块 文档；类别 docs；类型 documentation）：新增 GroupNorm+SiLU 融合、QK norm rope 等记录，以及开放 PR 监察清单，帮助开发者了解主线优化和进行中的工作。

关键符号：print_model_catalog, print_results_table

关键源码片段

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

核心 benchmark 脚本，修改了 model presets 和打印格式，新增 LTX-2.3 preset，调整表格宽度，使预设与 nightly 对齐。

```
# Table width constants for printed output alignment
CATALOG_TABLE_WIDTH = 105
RESULTS_TABLE_WIDTH = 105

# Model presets dict: keys are preset names used as --model argument
# Nightly-aligned presets come first, then skill-only extras.
# Each entry produces a sglang generate command.
MODELS = {
    # Example: flux1_dev_t2i_1024
    "flux": {
        "nightly_case_id": "flux1_dev_t2i_1024",
        "path": "black-forest-labs/FLUX.1-dev",
        "prompt": "A futuristic cyberpunk city at night, neon lights reflecting on wet streets",
        "extra_args": [
            "--width=1024", "--height=1024",
            "--num-inference-steps=50", "--guidance-scale=4.0",
            "--dit-layerwise-offload", "false",
        ],
    },
    # ... ( 其他预设节略 )
    # LTX-2: 更新为 2 GPU + CFG parallel 配置
    "ltx2": {
        "nightly_case_id": "ltx2_twostage_t2v",
        "path": "Lightricks/LTX-2",
        "prompt": "A cat and a dog baking a cake together in a kitchen.",
        "extra_args": [
            "--pipeline-class-name=LTX2TwoStagePipeline",
            "--width=768", "--height=512",
            "--num-frames=121",
            "--num-inference-steps=50", "--guidance-scale=4.0",
            "--num-gpus=2", "--enable-cfg-parallel",
        ],
    },
    # 新增 LTX-2.3 TI2V 两阶段 preset，与 nightly 配置 comparison_configs.json 对齐
    "ltx23-ti2v-two-stage": {
```

```
        "nightly_case_id": "ltx2.3_twostage_ti2v_2gpus",
        "path": "Lightricks/LTX-2.3",
        "prompt": "The cat starts walking slowly towards the camera.",
        "image_path": str(ASSET_DIR / "cat.png"),
        "extra_args": [
            "--pipeline-class-name=LTX2TwoStagePipeline",
            "--width=768", "--height=512",
            "--num-frames=121",
            "--num-inference-steps=50", "--guidance-scale=4.0",
            "--num-gpus=2",
        ],
    },
    # ... 更多 presets
}
```

评论区精华

review 中 gemini-code-assist[bot] 提出了 4 条关于表格分隔符长度的建议（建议统一使用 105 字符），认为不一致会降低可读性。这些建议未在提交历史中明确体现是否采纳（第三个 commit 提到 ' 格式问题 '，但具体未涵盖所有建议）。BBuf 未回复评论。

- 表格分隔符长度对齐 (style): 未在后续提交中完全修复，BBuf 未确认。

风险与影响

- 风险：低风险。主要变更集中在文档和脚本格式，未触及核心推理逻辑。但需要注意：LTX-2 预设参数变更（分辨率、GPU 数、seed 移除）将导致历史 benchmark 数据不可直接比较；文档中 Open PR Watchlist 需要持续维护，否则可能误导开发者。
- 影响：影响范围限于使用这些 skill 的开发者和 CI 流程。对系统无影响。团队需要维护 Open PR Watchlist 准确性。
- 风险标记：性能基准偏差，文档需持续维护

关联脉络

- PR #24219 [diffusion] CI: change ground truth repo: 都与 diffusion skill 同步相关，本 PR 也强调 benchmark 预设与 nightly CI 对齐。
- PR #23148 [codex] diffusion: enable group norm silu fuse by default: 该 PR 是 GroupNorm+SiLU 融合，在 existing-fast-paths.md 中被记录为已合并的工作。
- PR #23938 Optimize large GroupNorm SiLU apply: GroupNorm SiLU 大形状优化，也在 existing-fast-paths.md 中记录。