PR #25842 完整报告

sgl-project/sglang

[codex] Align diffusion skills with nightly Nvidia benchmarks

合并时间：2026-05-20 12:18

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25842

执行摘要

- 一句话：对齐 diffusion benchmark 预设至 Nvidia nightly 配置
- 推荐动作：建议阅读 bench_diffusion_denoise.py 中新增的 validate_nightly_alignment 函数，了解如何自动检查配置漂移。对于维护 diffusion benchmark 的团队，该 PR 提供了一个可持续对齐 nightly 配置的机制。

功能与动机

The diffusion skills had drifted from the current Nvidia nightly benchmark configuration. Some presets still encoded old/default sampling params, Wan commands used a shorthand instead of the explicit nightly size, and the LTX-2.3 TI2V preset missed the CI parallelism flag.

实现拆解

1. 在 bench_diffusion_denoise.py 中导入 shlex 并定义 NIGHTLY_CONFIG_PATH，指向 comparison_configs.json。
2. 定义 NIGHTLY_PRESET_ORDER 元组，严格反映 nightly 预设的顺序。
3. 从所有 nightly 对齐的 MODELS 条目中移除显式的 --num-inference-steps 和 --guidance-scale 参数，依赖运行时默认值；Wan 模型改用 --width=1280 --height=720 替代 --720p 快捷方式；LTX-2.3 增加 --cfg-parallel-size=2。
4. 新增 --validate-nightly-alignment 命令行选项，通过 _expected_nightly_cli_args 函数比对当前预设与 nightly config，若不一致则报错退出。
5. 同步更新三份文档文件（benchmark-and-profile.md、SKILL.md），说明移除显式采样参数的决策，并加入验证命令示例。

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 dependency-wiring；符号 _parse_cli_args, _normalize_cli_value, _expected_nightly_cli_args, validate_nightly_alignment）：核心基准测试脚本，大部分配置对齐和新验证逻辑在此文件实现。
- python/sglang/multimodal_gen/.claude/skills/sglang-diffusion-benchmark-profile/benchmark-and-profile.md（模块 文档；类别 docs；类型 documentation）：主要说明文档，同

步更新了配置对齐的说明和验证命令示例。

- python/sglang/multimodal_gen/.claude/skills/sglang-diffusion-performance/SKILL.md（模块 文档；类别 docs；类型 documentation）：性能优化技能文档，同步修改了命令示例以对齐 nightly 配置。
- python/sglang/multimodal_gen/.claude/skills/sglang-diffusion-benchmark-profile/SKILL.md（模块 文档；类别 docs；类型 documentation）：基准测试技能的主文档，简略提及新验证功能。

关键符号：_parse_cli_args, _normalize_cli_value, _expected_nightly_cli_args, validate_nightly_alignment

关键源码片段

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

核心基准测试脚本，大部分配置对齐和新验证逻辑在此文件实现。

```
import shlex # 用于安全拆分命令行参数

# Nightly 比较配置文件路径
NIGHTLY_CONFIG_PATH = (
    REPO_ROOT / "scripts" / "ci" / "utils" / "diffusion" / "comparison_configs.json"
)

# Nightly 预设顺序，与 comparison_configs.json 中的 case 顺序一致
NIGHTLY_PRESET_ORDER = (
    "flux",
    "flux2",
    "qwen",
    "qwen-edit",
    "zimage",
    "wan-t2v",
    "wan-ti2v",
    "ltx2",
    "ltx23-ti2v-two-stage",
    "wan-i2v",
)

# 模型配置字典（节选）
# 注意：nightly 对齐的 preset 不再显式指定 --num-inference-steps 和 --guidance-scale，
# 而是依赖 SGLang 运行时默认值，以匹配 comparison_configs.json 的省略语义。
MODELS = {
    # Nightly: 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",
            "--dit-layerwise-offload",
            "false",
            # 显式采样参数已移除，runtime 默认值为 50 steps, guidance 4.0
        ],
    },
    # Nightly: wan22_t2v_a14b_720p
    "wan-t2v": {
        "nightly_case_id": "wan22_t2v_a14b_720p",
        "path": "Wan-AI/Wan2.2-T2V-A14B-Diffusers",
        "prompt": "A cat and a dog baking a cake together in a kitchen.",
        "extra_args": [
            "--width=1280",
            "--height=720", # 替代原先的 --720p 快捷方式，显式使用 720p 尺寸
            "--num-frames=81",
            "--num-gpus=4",
            "--enable-cfg-parallel",
            "--ulysses-degree=2",
            "--text-encoder-cpu-offload",
            "--pin-cpu-memory",
            # steps 和 guidance 已移除，使用 runtime 默认值
        ],
    },
    # ... 其他模型类似
}
```

评论区精华

无审核评论讨论。

- 暂无高价值评论线程

风险与影响

- 风险：基准测试配置更改可能导致本地运行结果与预期不符，但验证命令可提前捕获差异。风险较低，主要在开发环境而非生产。同时依赖外部 comparison_configs.json 文件，若格式变化需同步更新解析逻辑。
- 影响：主要影响使用该 benchmark 脚本的开发者和 CI 流程。用户需更新脚本或依赖 nightly 配置。不会影响 SGLang 运行时核心功能。对于持续集成，降低了维护成本，因为配置漂移可被自动检测。
- 风险标记：配置漂移检测 , 依赖外部文件

关联脉络

- 暂无明显关联 PR