PR #24288 完整报告

sgl-project/sglang

[SKILL] Add diffusion benchmark presets for edit and Hunyuan3D models

合并时间：2026-05-05 08:18

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24288

执行摘要

- 一句话：新增 4 个 diffusion benchmark 预设（图像编辑和 3D 形状）
- 推荐动作：值得 benchmark 使用者和大模型 perf engineers 关注。PR 展示了如何通过实测数据（FireRed 1.0/1.1）做出多 GPU 策略选择（CFG parallel vs. Ulysses），以及如何用 config_overrides 实现模型级配置覆盖。但需注意文档与脚本的同步问题，建议采纳 review 中的修改建议。

功能与动机

需要系统化评测扩散模型在图像编辑和 3D 生成场景下的去噪延迟和峰值显存，为后续优化提供可复现的基准。PR Body 提供了 FireRed 在单 GPU 与双 GPU Ulysses/CFG parallel 下的对比数据，说明选择 CFG parallel 的决策依据。

实现拆解

1. 新增 benchmark 预设：在 bench_diffusion_denoise.py 的 MODELS 字典中添加 joyai-edit、firered-edit-1.0、firered-edit-1.1 和 hunyuan3d-shape 四个预设，每个预设包含模型路径、提示词、输入图片路径和额外参数。
2. 引入 config_overrides 机制：为支持 Hunyuan3D-2 的 paint_enable: False 配置，在 build_sglang_cmd 函数中新增 config_overrides 处理逻辑：将字典序列化为 JSON 文件并通过 --config 参数传递给 sglang generate。
3. 更新 GPU 需求函数：在 required_gpus_for_model 中添加 joyai-edit、firered-edit-1.0、firered-edit-1.1 三个预设返回 2 GPU 的需求。
4. 完善文档：在 benchmark-and-profile.md 中新增预设表格条目和手动命令示例；在 SKILL.md（性能文档）中添加图像编辑和 3D shape 的执行命令，并总结多 GPU 选择结论；在基准测试技能入口 SKILL.md 中更新参考列表。

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 configuration；符号 required_gpus_for_model, build_sglang_cmd）：核心 benchmark 脚本，新增 4 个模型预设和 config_overrides 处理逻辑，支撑 image-edit 和 3D shape 基准测试。
- python/sglang/multimodal_gen/.claude/skills/sglang-diffusion-benchmark-profile/benchmark-and-profile.md（模块 文档；类别 docs；类型 documentation）：benchmark 目录

文档，新增了 JoyAI/FireRed/Hunyuan3D 的手动命令示例和预设表项，帮助用户复现基准测试。

- python/sglang/multimodal_gen/.claude/skills/sglang-diffusion-performance/SKILL.md（模块 文档；类别 docs；类型 documentation）：性能文档，新增了图像编辑和 3D shape 的基准测试命令与参数建议，总结了多 GPU 选择结论。
- python/sglang/multimodal_gen/.claude/skills/sglang-diffusion-benchmark-profile/SKILL.md（模块 文档；类别 docs；类型 documentation）：基准测试技能入口文档，更新了主参考列表，新增对图像编辑和 3D shape presets 的引用。

关键符号：required_gpus_for_model, build_sglang_cmd

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，新增 4 个模型预设和 config_overrides 处理逻辑，支撑 image-edit 和 3D shape 基准测试。

```
# MODELS 字典新增预设 — 保持与 benchmark-and-profile.md 同步
MODELS = {
    # 16. Skill-only extra preset: JoyAI Image Edit
    "joyai-edit": {
        "path": "jdopensource/JoyAI-Image-Edit-Diffusers",
        "prompt": "Make the cat wear a red hat",
        "image_path": str(ASSET_DIR / "cat.png"),
        "extra_args": [
            "--width=1024", "--height=1024", "--num-inference-steps=40",
            "--guidance-scale=4.0", "--dit-layerwise-offload false",
            "--dit-cpu-offload false", "--num-gpus=2",
            "--enable-cfg-parallel", "--ulysses-degree=1",
        ],
    },
    # 17. Skill-only extra preset: FireRed Image Edit 1.0
    "firered-edit-1.0": {
        "path": "FireRedTeam/FireRed-Image-Edit-1.0",
        "prompt": "Make the cat wear a red hat",
        "image_path": str(ASSET_DIR / "cat.png"),
        "extra_args": [
            "--width=1024", "--height=1024", "--num-inference-steps=40",
            "--guidance-scale=4.0", "--dit-layerwise-offload false",
            "--dit-cpu-offload false", "--num-gpus=2",
            "--enable-cfg-parallel", "--ulysses-degree=1",
        ],
    },
    # 18. Skill-only extra preset: FireRed Image Edit 1.1
    "firered-edit-1.1": {
        "path": "FireRedTeam/FireRed-Image-Edit-1.1",
        "prompt": "Make the cat wear a red hat",
```

```
        "image_path": str(ASSET_DIR / "cat.png"),
        "extra_args": [
            "--width=1024", "--height=1024", "--num-inference-steps=40",
            "--guidance-scale=4.0", "--dit-layerwise-offload false",
            "--dit-cpu-offload false", "--num-gpus=2",
            "--enable-cfg-parallel", "--ulysses-degree=1",
        ],
    },
    # 19. Skill-only extra preset: Hunyuan3D Shape Generation
    "hunyuan3d-shape": {
        "path": "tencent/Hunyuan3D-2",
        "prompt": "generate 3d mesh",
        "image_path": str(ASSET_DIR / "cat.png"),
        "config_overrides": {"paint_enable": False}, # 关闭上色阶段，聚焦去噪
        "extra_args": [
            "--num-inference-steps=50", "--guidance-scale=5.0",
            "--dit-layerwise-offload false", "--dit-cpu-offload false",
        ],
    },
}

# 在 build_sglang_cmd 中处理 config_overrides
if "config_overrides" in cfg:
    # 注意：当前写入 ASSET_DIR 的子目录，可能因只读失败；建议改用输出目录
    config_dir = ensure_dir(ASSET_DIR / "generated_configs")
    config_path = config_dir / f"{model_key}.json"
    with open(config_path, "w") as f:
        json.dump(cfg["config_overrides"], f)
    cmd.append(f"--config={config_path}")
```

评论区精华

在 Code Review 中，gemini-code-assist[bot] 提出两个关键问题：

1)config_overrides写入的目录应使用benchmark输出目录而非输入资产目录（ASSET_DIR），以避免只读环境问题； 2) benchmark-and-profile.md 中的 Hunyuan3D 手动命令示例未包含 --config 参数以禁用 paint stage，与脚本预设不同步。目前这两个问题在 PR 中未明确解决，需后续跟进。

- config 文件写入路径应使用输出目录而非 ASSET_DIR (design): 建议未明确确认是否采纳，从最终代码看可能保留原实现，存在只读风险。
- Hunyuan3D 手动命令示例缺少关闭上色阶段的配置 (documentation): 未明确采纳，文档示例仍缺少该配置，可能导致用户复现结果不一致。

风险与影响

- 风险：

1. 配置路径可写性：当前 config_overrides 生成的 JSON 文件写入 ASSET_DIR/generated_configs，若 ASSET_DIR 为只读则会导致脚本失败。
2. 文档同步偏离：benchmark-and-profile.md 中 Hunyuan3D 手动示例未体现 paint_enable: False 配置，用户可能复现出包含 paint 阶段的完整流水线，导致去噪延迟偏高且不可比。
3. 多 GPU 策略固化：预设强制使用 2-GPU CFG parallel，但未考虑其他硬件环境（如 A100 或不同显存），直接复用预设可能因显存不足而 OOM。 - 影响：影响范围：扩散模型 benchmark 流程和性能文档使用者。影响程度：中低。新增的预设降低了 image-edit 和 3D shape 基准测试的配置门槛，但核心脚本和文档均位于 .claude/skills 目录，不是主代码路径，不会影响 SRT 运行时。对团队而言，这些预设为后续 diffusion 优化提供了标准化评估手段。

- 风险标记：配置路径只读风险，文档与脚本预设不同步

关联脉络

- PR #23200 [Diffusion] Enable channels-last 3D VAE convs by default: 同为 diffusion 性能优化 PR，本 PR 的 benchmark 预设可用于验证 channels-last 优化的效果。
- PR #24366 [diffusion] Use direct all-to-all for USP collectives: 同为 diffusion 通信优化 PR，本 PR 新增的 FireRed 预设（使用 CFG parallel）可评估不同通信策略的差异。