PR #25483 完整报告

sgl-project/sglang

[codex] Update Wan2.2 ModelOpt CI checkpoints

合并时间：2026-05-20 09:05

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25483

执行摘要

- 一句话：更新 Wan2.2 ModelOpt CI 检查点为 NVIDIA 官方版本
- 推荐动作：建议阅读该 PR 以了解 SGLang 对 ModelOpt NVFP4 布局的处理方式，特别是 swap_weight_nibbles 的默认值选择逻辑和 FLUX.1 特殊分支。对维护 Blackwell 量化的开发者具有参考价值。

功能与动机

PR body：替换 Wan2.2 ModelOpt FP8/NVFP4 CI 用例为 NVIDIA 官方完整 Diffusers ModelOpt 发布版本：nvidia/Wan2.2-T2V-A14B-Diffusers-FP8 和 nvidia/Wan2.2-T2V-A14B-Diffusers-NVFP4。从 CI 用例定义中移除旧的 lmsys/*-sglang-transformer 覆盖层。

实现拆解

1. 调整 NVFP4 配置默认值：将 ModelOptFp4Config.__init__ 中 swap_weight_nibbles 默认值从 True 改为 False（modelopt_quant.py）。在 from_config 中，当未显式设置时回退到 checkpoint_uses_packed_qkv 的值。同步修改 _merge_modelopt_fp4_configs（transformer_load_utils.py）。
2. 修复 FLUX.1 NVFP4 CUDNN scale 布局：在 process_weights_after_loading 中增加 FLUX.1 特定分支，检测 prefix 是否以 transformer_blocks. 或 single_transformer_blocks. 开头且非 packed QKV，强制使用 CUTLASS/TMA scale 布局（modelopt_quant.py）。
3. 注册 NVIDIA 官方 checkpoint：在 registry.py 中添加 nvidia/Wan2.2-T2V-A14B-Diffusers-NVFP4 路径。
4. 更新 CI 测试用例：在 gpu_cases.py 中将 Wan2.2 FP8/NVFP4 用例的 model_path 改为官方 repo，移除 --transformer-path，设置 run_consistency_check=True；调整 NVFP4 环境变量为 trtllm 后端。在 testcase_configs.py 中添加对应常量。
5. 启用一致性检查：为六个 ModelOpt 用例启用 run_consistency_check=True，新增 consistency_threshold.json 定义阈值。
6. 更新文档：在 quantization.mdx 中更新支持矩阵和 CLI 示例，移除 --transformer-path 引用。

关键文件：

- python/sglang/multimodal_gen/runtime/layers/quantization/modelopt_quant.py（模块 量化配置；类别 source；类型 data-contract；符号 ModelOptFp4Config.init, ModelOptFp4Config.from_config, ModelOptFp4LinearMethod.process_weights_after_loading）：核心量化配置类，调整 swap_weight_nibbles 默认值并修复 FLUX.1 CUDNN scale 布局
- python/sglang/multimodal_gen/runtime/loader/transformer_load_utils.py（模块 加载工具；类别 source；类型 core-logic；符号 _merge_modelopt_fp4_configs）：合并 NVFP4 配置的核心逻辑，确保推断配置保留 swap_weight_nibbles 等关键字段
- python/sglang/multimodal_gen/tools/build_modelopt_nvfp4_transformer.py（模块 构建工具；类别 source；类型 data-contract；符号 build_modelopt_nvfp4_transformer, _parse_args）：构建 NVFP4 transformer 的工具，默认 swap_weight_nibbles 统一改为 False
- python/sglang/multimodal_gen/registry.py（模块 注册表；类别 source；类型 core-logic；符号 _register_configs）：注册 NVIDIA 官方 NVFP4 checkpoint 路径
- python/sglang/multimodal_gen/test/server/gpu_cases.py（模块 GPU 测试；类别 test；类型 test-coverage）：更新 CI 测试用例指向新 checkpoint，启用一致性检查
- docs_new/docs/sglang-diffusion/quantization.mdx（模块 文档；类别 other；类型 core-logic）：更新文档中的 checkpoint 表格和 CLI 示例
- python/sglang/multimodal_gen/test/server/testcase_configs.py（模块 测试配置；类别 test；类型 test-coverage）：新增 NVIDIA 官方模型路径常量
- python/sglang/multimodal_gen/test/server/consistency_threshold.json（模块 一致性阈值；类别 test；类型 test-coverage）：新增一致性检查阈值定义
- python/sglang/multimodal_gen/test/test_utils.py（模块 测试工具；类别 test；类型 test-coverage）：测试工具调整（可能涉及一致性检查工具）

关键符号：ModelOptFp4Config.init, ModelOptFp4Config.from_config, ModelOptFp4LinearMethod.process_weights_after_loading, _merge_modelopt_fp4_configs, build_modelopt_nvfp4_transformer, _register_configs

关键源码片段

python/sglang/multimodal_gen/runtime/loader/transformer_load_utils.py

合并 NVFP4 配置的核心逻辑，确保推断配置保留 swap_weight_nibbles 等关键字段

```
def _merge_modelopt_fp4_configs(
    existing_config: Optional[ModelOptFp4Config],
    inferred_config: Optional[ModelOptFp4Config],
) -> Optional[ModelOptFp4Config]:
    """
    合并来自 config.json 的现有配置和从 safetensors shards 推断的配置。
    优先使用推断的 exclude_modules（更准确），但保留 repo 级别的
    swap_weight_nibbles 和 checkpoint_uses_packed_qkv 设置。
    """
    if inferred_config is None:
```

```
        return existing_config
    if _get_quant_config_name(inferred_config) != "modelopt_fp4":
        return existing_config or inferred_config
    if existing_config is None:
        return inferred_config
    if _get_quant_config_name(existing_config) != "modelopt_fp4":
        return existing_config

    existing_excludes = getattr(existing_config, "exclude_modules", []) or []
    inferred_excludes = getattr(inferred_config, "exclude_modules", []) or []
    if inferred_excludes != existing_excludes:
        logger.warning(
            "Overriding ModelOpt NVFP4 exclude_modules from config.json with "
            "safetensors-inferred layout (%d -> %d entries).",
            len(existing_excludes),
            len(inferred_excludes),
        )

    inferred_config.packed_modules_mapping = getattr(
        existing_config, "packed_modules_mapping", {}
    )
    # 保留推断配置中的 checkpoint_uses_packed_qkv，若未设置则从 existing_config 继承
    inferred_config.checkpoint_uses_packed_qkv = getattr(
        inferred_config, "checkpoint_uses_packed_qkv", False
    ) or getattr(existing_config, "checkpoint_uses_packed_qkv", False)
    # swap_weight_nibbles 优先取推断配置，否则取 existing_config，均默认为 False
    inferred_config.swap_weight_nibbles = getattr(
        inferred_config, "swap_weight_nibbles", False
    ) or getattr(existing_config, "swap_weight_nibbles", False)
    if getattr(inferred_config, "group_size", None) is None:
        inferred_config.group_size = getattr(existing_config, "group_size", None)

    return inferred_config
```

评论区精华

1. 文档表格格式争议：gemini-code-assist 建议将 Base Model 列使用原始未量化模型名称以求统一，PR 未采纳。
2. 文档迁移要求：zijiexia 要求将变更移到 docs_new 目录，BBuf 确认执行。
3. NVFP4 绿屏根因分析：BBuf 发现 NVIDIA 官方 NVFP4 checkpoint 缺少 swap_weight_nibbles 配置，导致 nibble 交换错误产生绿屏视频，通过默认改为 False 修复。
4. FLUX.1 CUDNN 布局移植：从 #25527 移植 FLUX.1 特定的 CUDNN scale 布局修复。
5. 一致性检查启用：BBuf 为六个 ModelOpt CI 用例启用输出一致性检查。

- 文档表格中 Base Model 列内容 (style): 未采纳，PR 保持使用量化 repo ID。
- 文档迁移到 docs_new 目录 (documentation): BBuf 确认并将更改迁移到 docs_new 目录。

- NVIDIA Wan2.2 NVFP4 nibble 布局修复 (bugfix): 通过 commit 9fac9a0 和 ddfffc9 修复。
- FLUX.1 NVFP4 CUDNN scale 布局移植 (bugfix): 通过 commit 39b33a1 修复。
- 启用 ModelOpt 扩散一致性检查 (testing): 通过 commit 5ac8232 启用。

风险与影响

- 风险:
  - swap_weight_nibbles 默认值变更：可能影响其他未显式设置该值的 NVFP4 checkpoint，但 FLUX.2 等使用 checkpoint_uses_packed_qkv=True 保持了正确行为。
- CUDNN 路径性能差异：FLUX.1 强制 CUTLASS/TMA 布局可能带来性能变化，但保证了正确性。
- CI 一致性阈值：新增阈值需持续观察以避免假阳性。
- 影响:
  - 用户：可直接使用 --model-path nvidia/... 加载官方量化 checkpoint，无需额外 transformer 覆盖层。
- CI：FP8 用例在 H100 shard 运行，NVFP4 在 B200 shard。
- 系统：量化配置默认值改变，但已验证兼容现有发布的 checkpoint。
- 风险标记：swap_weight_nibbles 默认变更，CUDNN 布局路径差异，一致性阈值待观察

关联脉络

- PR #25527 FLUX.1 NVFP4 CUDNN scale-layout fix: 本 PR 合并了此修复以修正 FLUX.1 NVFP4 CUDNN scale 布局。