PR #28928 完整报告

sgl-project/sglang

[diffusion] Add Qwen-Image ModelOpt NVFP4 support

合并时间：2026-06-25 22:56

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28928

执行摘要 该 PR 为 SGLang Diffusion 的 Qwen-Image 系列模型添加了 ModelOpt NVFP4 量化加载支持，基于 NVIDIA ModelOpt 导出的 checkpoint，在 Blackwell 硬件上显存降低约 63%，并保持图片质量。变更包括注册 NVFP4 模型路径、调制层名称规范化、添加一致性测试及更新文档。

功能与动机 PR body 指出：“基于 ModelOpt PR #1706 导出的 NVFP4 checkpoint，为 SGLang Diffusion 增加原生加载支持”。动机是使 Qwen-Image、Qwen-Image-2512、Qwen-Image-Edit 等模型能够利用 NVFP4 的低比特精度，在 B200/B300 上高效推理。

实现拆解

1. 模型注册：在 registry.py 中为 Qwen-Image 增加 nvidia/Qwen-Image-NVFP4 路径，使 NVFP4 完整 repo 可通过 --model-path 直接加载。
2. 名称规范化：quantization_utils.py 新增 _canonicalize_modulation_exclude，将序列化权重中的 .img_mod.1/.txt_mod.1 映射为运行时前缀 .img_mod/.txt_mod，确保量化排除列表正确。
3. 配置集成：在 _build_nvfp4_config_from_safetensors_files 中应用上述规范化，使得从 safetensors metadata 推断的排除模块名称可匹配运行时层。
4. CI 测试：在 gpu_cases.py 为 qwen_image_2512_modelopt_nvfp4_t2i 注册 B200 50 步一致性测试，并更新 perf_baselines.json 和 consistency_threshold.json。
5. 部署文档：qwen-image-deployment.jsx 增加硬件平台和精度选择器，quantization.mdx 更新 NVFP4 支持模型列表和用法说明。

关键源码片段 下面是调制层名称规范化的核心函数实现： def_canonicalize_modulation_exclude(module_name: str) -> str: """映射序列化调制权重的父名称到运行时间线性前缀. Qwen-Image 将调制投影包装在 ``nn.Sequential(SiLU, Linear)`` 中，因此权重序列化为 ``...img_mod.1.weight``，而运行时的 ReplicatedLinear 以 ``...img_mod`` 作为量化/排除前缀. 去掉 Sequential 索引使 safetensors 推断的 BF16 排除条目 能与线性层匹配（与 ModelOpt FP8 转换器保持一致，后者将 ``.img_mod.1`` / ``.txt_mod.1`` 规范化为 ``.img_mod`` / ``.txt_mod``）. 对其他模块名称无操作. """ if module_name.endswith(('.img_mod.1', '.txt_mod.1')): return module_name.removesuffix('.1') return module_name

评论区精华

- 规范化函数简化：gemini-code-assist[bot] 建议使用 endswith(tuple) + removesuffix，BBuf 采纳并提交简化版本。

- 文档更新建议：mickqian 询问是否更新 cookbook，BBuf 确认已更新。

风险与影响

- 风险：NVFP4 路径依赖特定 ModelOpt 导出格式；仅 Blackwell 硬件可用；仅一个 case 有 CI 覆盖；性能估计时间可能不足。
- 影响：用户可直接加载 lmsys/* 的 NVFP4 checkpoint，显存减半；CI 增加约 2 分钟。

关联脉络 该 PR 基于 #28557 的内定草稿，并使用 ModelOpt PR #1706 的导出工具。是扩散模型 NVFP4 支持系列的重要一步。

实现拆解

docs_new/src/snippets/diffusion/qwen-image-deployment.jsx

部署交互组件：增加硬件平台（B200/B300/H200/H100）和精度（BF16/NVFP4）选择，根据配置自动切换模型路径。

```
const config = {
    modelFamily: 'Qwen-Image',
    options: {
        hardware: {
            items: [
                // 默认硬件改为 B200
                { id: 'b200', label: 'B200', default: true },
                { id: 'b300', label: 'B300', default: false },
                // ... 其他硬件
            ]
        },
        precision: {
            name: 'precision',
            title: 'Precision',
            items: [
                { id: 'bf16', label: 'BF16', default: true },
                {
                    id: 'nvfp4',
                    label: 'NVFP4',
                    default: false,
                    disabledWhen: (values) => !['b200', 'b300'].includes(values.hardware),
                    disabledReason: 'ModelOpt NVFP4 requires Blackwell hardware such as B200 or
                    B300'
                }
            ]
        }
    },
    generateCommand: function(values) {
        const isBlackwell = ['b200', 'b300'].includes(values.hardware);
        const isNvfp4 = values.precision === 'nvfp4' && isBlackwell;
        // NVFP4 使用 lmsys 发布的量化 checkpoint
        const modelPath = isNvfp4
            ? 'lmsys/qwen-image-2512-modelopt-nvfp4-sglang'
```

```
            : 'Qwen/Qwen-Image';
        return `sglang serve --model-path ${modelPath} ...`;
    }
};

// 切换硬件时自动重置精度
const handleRadioChange = (optionName, value) => {
    setValues((prev) => {
        const next = { ...prev, [optionName]: value };
        if (optionName === 'hardware' && !['b200', 'b300'].includes(value) && next.precision ===
        'nvfp4') {
            next.precision = 'bf16';
        }
        return next;
    });
};
```