PR #23625 完整报告

sgl-project/sglang

Flux2 nvfp4 quantization correctness on Blackwell (B200)

合并时间：2026-05-02 09:57

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23625

执行摘要

- 一句话：修复 FLUX.2 NVFP4 在 B200 上的量化正确性
- 推荐动作：值得精读。关注 process_weights_after_loading 的条件化设计、per-GEMM z-score 调试方法。建议未来建立可配置命名映射机制。

功能与动机

在 B200 硬件上，FLUX.2-dev-NVFP4 量化模型输出全白图像（均值≈252，标准差≈2），而 BF16 版本输出正常。经调查，上游 main 分支存在三个正确性 bug，共同导致 NVFP4 路径失效。详见 PR #23625 body。

实现拆解

1. 修复 Input Scale 缺失：在 flux_2_nvfp4.py 中为 16 个 FP4 量化 txt_mlp 层补全 input_scale 参数。
2. 条件化 TMA Scale 排列：在 modelopt_quant.py 的 process_weights_after_loading 中根据后端类型决定是否进行 blockwise interleave。
3. 修复加载器回退：在 quantization_utils.py 修正排除模块映射，在 mlp.py 和 wanvideo.py 中统一前缀命名。
4. 增强检查点恢复：在 utils.py 新增 _try_redownload_missing_shards 函数，自动修复不完整检查点。
5. 兼容性修复与 CI 恢复：在 component_loader.py 增加 RobertaProcessing 回退，恢复 B200 CI。

关键文件：

- python/sglang/multimodal_gen/runtime/loader/utils.py（模块 加载工具；类别 source；类型 dependency-wiring；符号 _try_redownload_missing_shards）：核心修复：添加检查点完整性校验和自动修复，防止缓存不完整导致的故障
- python/sglang/multimodal_gen/runtime/layers/quantization/modelopt_quant.py（模块 量化层；类别 source；类型 data-contract；符号 process_weights_after_loading, create_weights）：修复 TMA scale 排列条件化，确保 cuDNN 后端获得正确的 row-major scales

- python/sglang/multimodal_gen/runtime/loader/component_loaders/component_loader.py（模块 组件加载；类别 source；类型 core-logic；符号 load_customized）：修复 tokenizers>=0.21 的 RobertaProcessing 兼容错误，恢复 B200 CI
- python/sglang/multimodal_gen/runtime/models/dits/wanvideo.py（模块 模型定义；类别 source；类型 data-contract；符号 WanSelfAttention, WanTransformerBlock, WanTransformerBlock_VSA）：修正参数前缀命名以匹配排除模块规则，确保 NVFP4 加载路径生效
- python/sglang/multimodal_gen/runtime/utils/quantization_utils.py（模块 量化工具；类别 source；类型 core-logic；符号 _build_nvfp4_config_from_safetensors_files）：修复排除模块映射逻辑中的 .weight 后缀问题
- python/sglang/multimodal_gen/runtime/layers/mlp.py（模块 MLP 层；类别 source；类型 core-logic；符号 MLP）：修正 MLP 线性层前缀命名以匹配模型键，支持 NVFP4 排除

关键符号：_try_redownload_missing_shards, _list_safetensors_files, process_weights_after_loading, create_weights, _build_nvfp4_config_from_safetensors_files, load_customized

关键源码片段

python/sglang/multimodal_gen/runtime/layers/quantization/modelopt_quant.py

修复 TMA scale 排列条件化，确保 cuDNN 后端获得正确的 row-major scales

```
def process_weights_after_loading(self, layer: torch.nn.Module) -> None:
    # ... 前面的 alpha, input_scale_inv 计算
    scales = layer.weight_scale
    scale_ndim = scales.ndim
    if scale_ndim == 2:
        scales = scales.unsqueeze(0)
    assert scales.ndim == 3
    B, M, K = scales.shape
    M_padded = round_up(M, 128)
    K_padded = round_up(K, 4)
    padded_scales = torch.zeros((B, M_padded, K_padded), dtype=scales.dtype)
    padded_scales[:B, :M, :K] = scales

    # 关键变更：仅在 CUTLASS 路径下应用 TMA 排列
    _, flashinfer_backend = _get_fp4_gemm_op()
    if flashinfer_backend is None:
        # CUTLASS (sgl_kernel) 路径：blockwise interleave 适应 TMA 布局
        padded_scales = padded_scales.reshape(
            B, M_padded // 128, 4, 32, K_padded // 4, 4
        )
        padded_scales = padded_scales.permute(0, 1, 4, 3, 2, 5)

    padded_scales = padded_scales.contiguous().cuda()
    padded_scales = (
```

```
            padded_scales.reshape(M_padded, K_padded)
            if scale_ndim == 2
            else padded_scales.reshape(B, M_padded, K_padded)
        )
        copy_or_rebind_param(layer, 'weight_scale_interleaved', padded_scales)
```

评论区精华

在 Review 中，BBuf 指出 swap_weight_nibbles 在 from_config 中已默认 True，作者移除显式设置。BBuf 询问是否仅改 modelopt_quant.py 即可修复，作者确认需三个修复同时应用。OrangeRedeng 质疑 mlp.py 和 wanvideo.py 的 prefix 更名影响 Wan 模型，该兼容性问题未彻底解决。

- swap_weight_nibbles 默认值冗余 (design): 作者移除该显式设置，E2E 行为无变化。
- 修复范围是否仅需 modelopt_quant.py (correctness): 作者验证后确认需要三个修复同时应用。
- prefix 命名兼容性影响 Wan 模型 (question): 作者表示更名是为了匹配排除模块命名，当前 PR 已合并，后续需多架构适配。

风险与影响

- 风险： 1) prefix 修改可能影响其他模型（如 Wan），已由 CI 失败证实。 2) 自动修复依赖 hf_hub_download 网络，离线环境失败。 3) 仅 B200 验证，Hopper/Ada 未回归。 4) 无新增单元测试覆盖自动修复路径。
  - 影响：正面：FLUX.2-dev-NVFP4 在 B200 上恢复正常图像质量，检查点自动修复提高鲁棒性。负面：Wan 等模型可能因 prefix 变更而量化加载失败，需后续适配。
  - 风险标记：核心路径变更（量化路径），跨模型兼容性（Wan prefix 影响），无新增单测覆盖自动修复，仅验证 B200 硬件，离线环境自动修复依赖 HuggingFace Hub 网络

关联脉络

- PR #23155 [Diffusion] Add Qwen Image ModelOpt FP8 support: 同为 diffusion 量化支持，可能共享加载和量化基础设施
- PR #24315 [diffusion] chore: disable VAE cpu offload by default: 同属 diffusion 模块性能优化，修改了 server_args 等公共配置
- PR #18764 [diffusion] Add dynamic batching v0: 同属 diffusion 模块功能扩展，涉及模型加载和调度逻辑