PR #41616 完整报告

vllm-project/vllm

Test nemotron nano-v2 and nemotron nano-v3 separately, disable super-omni redundant tests

合并时间：2026-05-04 21:31

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/41616

执行摘要

- 一句话：分离 Nemotron nano-v2/v3 测试，禁用冗余测试
- 推荐动作：该 PR 虽然只修改了测试配置，但体现了测试策略的微调：将不同变体映射到各自模型以触发差异特性。值得关注的决策点是移除 V2 的 vision_config override，未来若出现 OOM 问题可快速恢复。建议 CI 运行后观察 V2 测试是否稳定。

功能与动机

根据 PR body 中的说明：

- Test nvidia/NVIDIA-Nemotron-Nano-12B-v2-VL-BF16 as nano-vl-v2, so we hit static resolution paths. * Test nvidia/Nemotron-3-Nano-Omni-30B-A3B-Reasoning-BF16 proper, so we also hit audio processing paths (which nano-vl-v2 doesn't support) * Disable (for now) NemotronH_Super_Omni_Reasoning_V3: It is just running nano-vl-v3, so we don't gain anything from the duplication, it's just wasting resources.

实现拆解

1. 修改 NemotronH_Nano_VL_V2 条目：将 hf_overrides 中的 vision_config 移除，仅保留 text_config 的 override，简化配置，但仍使用原模型 ID nvidia/NVIDIA-Nemotron-Nano-12B-v2-VL-BF16。
2. 修改 NemotronH_Nano_Omni_Reasoning_V3 条目：将其模型 ID 从原本指向 V2 的 nvidia/NVIDIA-Nemotron-Nano-12B-v2-VL-BF16 改为实际的 nvidia/Nemotron-3-Nano-Omni-30B-A3B-Reasoning-BF16，并保留完整的 vision_config 和 text_config override，包括新增 video_maintain_aspect_ratio=False 以绕过处理器 Bug。
3. 修改 NemotronH_Super_Omni_Reasoning_V3 条目：将其模型 ID 指向同样的 Nano Omni 模型，但设置 is_available_online=False，从而在测试中标记为不可在线获取，等效于禁用该测试。
4. 删除冗余注释和配置：清理了先前错误地认为 V3 是 V2 别名的注释，并移除重复的 vision_config 定义。

关键文件：

- tests/models/registry.py（模块 测试注册表；类别 test；类型 test-coverage）：唯一被修改的文件，集中体现了所有测试注册表条目的调整：分离 V2/V3、禁用冗余 Super Omni 测试。

关键符号：未识别

关键源码片段

tests/models/registry.py

唯一被修改的文件，集中体现了所有测试注册表条目的调整：分离 V2/V3、禁用冗余 Super Omni 测试。

```
# 修改后的 registry 条目片段

"NemotronH_Nano_VL_V2": _HfExamplesInfo(
    "nvidia/NVIDIA-Nemotron-Nano-12B-v2-VL-BF16",
    max_model_len=4096,
    use_original_num_layers=True,
    hf_overrides={
        # 只保留 text_config，移除了 vision_config
        "text_config": {"num_hidden_layers": 2, "hybrid_override_pattern": "M*"},
    },
    trust_remote_code=True,
),

"NemotronH_Nano_Omni_Reasoning_V3": _HfExamplesInfo(
    "nvidia/Nemotron-3-Nano-Omni-30B-A3B-Reasoning-BF16", # 改为实际 V3 模型
    max_model_len=4096,
    use_original_num_layers=True,
    hf_overrides={
        "vision_config": PretrainedConfig(
            args={
                "min_num_patches": 1,
                "max_num_patches": 12,
                "model": "vit_huge_patch16_224",
            },
            video_temporal_patch_size=2,
            # 绕过处理器 Bug：官方 config.json 中为 true，但测试中导致异常
            video_maintain_aspect_ratio=False,
        ),
        "text_config": {"num_hidden_layers": 2, "hybrid_override_pattern": "M*"},
    },
    trust_remote_code=True,
),

"NemotronH_Super_Omni_Reasoning_V3": _HfExamplesInfo(
    "nvidia/Nemotron-3-Nano-Omni-30B-A3B-Reasoning-BF16",
    is_available_online=False # 标记为不可在线获取，从而禁用测试
),
```

评论区精华

唯一的实质性 review 评论来自 gemini-code-assist[bot]，指出为 NemotronH_Nano_VL_V2 移除 vision_config override（之前限制了补丁数量以降低内存）可能导致 CI 中出现 OOM 问题，并失去动态分辨率和 conv3d 特征的测试覆盖。但该 PR 最终被 tomeras91 批准，说明团队认为这一 trade-off 可接受，可能因为 V2 测试本身运行在较小模型上且资源足够。

- 移除 vision_config override 导致 OOM 风险 (correctness): PR 仍被批准，团队默许当前简化；可能因 V2 测试资源足够，或 OOM 问题可通过其他方式缓解。

风险与影响

- 风险：
  1. OOM 风险：移除了 NemotronH_Nano_VL_V2 的 vision_config override（原有限制 max_num_patches 等），可能导致 CI 中内存不足，尤其是当模型加载完整视觉塔时。
  2. 测试覆盖损失：之前通过 V2 测试覆盖的动态分辨率和 conv3d 逻辑现在可能未被充分测试，因为 V3 条目中虽然保留了这些配置，但 V3 是不同模型，可能不完全覆盖相同特性。
  3. 依赖可用性：Nemotron-3-Nano-Omni-30B-A3B-Reasoning-BF16 模型可能并非总是在线可用，若模型无法下载会导致测试失败。 - 影响：直接影响范围仅限于 NemotronH 系列的测试注册表条目，不涉及核心逻辑。主要影响是测试效率提升（减少冗余）和测试路径的明确分离；潜在负面影响是可能引入 OOM 失败或覆盖缺口。团队已 approve，表明风险可控。 - 风险标记：测试覆盖减少，潜在 OOM 风险

关联脉络

- 暂无明显关联 PR