PR #21544 完整报告

sgl-project/sglang

[diffusion][mova]Enhance cfg parallel for mova and update CI configuration

合并时间：2026-05-26 21:59

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21544

执行摘要

- 一句话：MOVA 模型 CFG 并行与 CI 增强
- 推荐动作：建议精读 mova.py 中 parallelism_type 和 _cfg_combine 的设计权衡，尤其是 PR 描述与实现的差异。提取辅助方法的设计值得参考，但需注意串行分支未同步重构的遗留问题。测试用例的 CI 补充方式可作为模板。

功能与动机

提升 MOVA 视频生成模型的推理吞吐量，通过跨 GPU 分布式计算 CFG 的正负预测，避免单 GPU 串行执行瓶颈。PR 描述中明确提到 "split positive/negative predictions across multiple GPUs for improved throughput"。

实现拆解

1. 去除 CFG_PARALLEL 枚举返回值：在 mova.py 的 parallelism_type 属性中，移除了 if enable_cfg_parallel: return StageParallelismType.CFG_PARALLEL 逻辑，改为始终返回 REPLICATED（与 PR 描述矛盾，但符合实际调度需求）。
2. 设备处理优化：inference_single_step 中使用 get_local_torch_device() 替代从 latents 推断 device，增强跨 GPU 兼容性。
3. 新增 CFG 组合辅助方法：提取 _cfg_combine 私有方法，统一处理视觉和音频噪声预测的组合与 guidance rescale，减少重复代码。
4. 测试用例扩展：在 testcase_configs.py 中添加 MOVA 720p (1 GPU) 和 MOVA 360p CFG parallel (2 GPU) 的测试配置。

关键文件：

- python/sglang/multimodal_gen/runtime/pipelines_core/stages/model_specific_stages/mova.py（模块 扩散模型；类别 source；类型 core-logic；符号 parallelism_type, inference_single_step, _cfg_combine, forward）：核心变更文件：修改 parallelism_type 属性、新增 _cfg_combine 辅助方法、优化设备处理逻辑。
- python/sglang/multimodal_gen/test/server/testcase_configs.py（模块 测试；类别 test；类型 test-coverage；符号 DiffusionTestCase）：新增 MOVA 720p 和 CFG parallel 测试配置，确保 CI 覆盖。

关键符号：parallelism_type, inference_single_step, _cfg_combine, forward

关键源码片段

python/sglang/multimodal_gen/runtime/pipelines_core/stages/model_specific_stages/mova.py

核心变更文件：修改 parallelism_type 属性、新增 _cfg_combine 辅助方法、优化设备处理逻辑。

```
# 关键片段：parallelism_type 属性变更与 _cfg_combine 辅助方法

@property
def parallelism_type(self) -> StageParallelismType:
    # 注意：PR 描述称返回 CFG_PARALLEL，但实际始终返回 REPLICATED
    # 需确认调度器是否依赖此标志
    return StageParallelismType.REPLICATED


def _cfg_combine(self, pos_out, neg_out, guidance_scale, cfg_rank, enable_cfg_parallel):
    # 将正负预测组合，消除与串行分支的重复
    if enable_cfg_parallel and cfg_rank == 1:
        # 在 cfg_rank 1 上，neg 来自其他 GPU
        return pos_out + guidance_scale * (pos_out - neg_out)
    else:
        return pos_out


def forward(self, batch: Req, server_args: ServerArgs) -> Req:
    # ... 原有逻辑 ...
    # 调用 _cfg_combine 代替内联组合
    visual_noise_pred = self._cfg_combine(
        pos[0], neg[0], batch.guidance_scale, cfg_rank, enable_cfg_parallel
    )
    audio_noise_pred = self._cfg_combine(
        pos[1], neg[1], batch.guidance_scale, cfg_rank, enable_cfg_parallel
    )
    if batch.guidance_rescale > 0.0:
        # guidance rescale 逻辑
        pass
```

python/sglang/multimodal_gen/test/server/testcase_configs.py

新增 MOVA 720p 和 CFG parallel 测试配置，确保 CI 覆盖。

```
# 测试配置片段
DiffusionTestCase(
    "MOVA-720p",
    model="MOVA-720p",
    num_gpus=1, # 仅验证路径可用
    run_perf_check=False,
),
DiffusionTestCase(
```

```
    "MOVA-360p-CFG-Parallel",
    model="MOVA-360p",
    num_gpus=2, # 双 GPU CFG 并行
    enable_cfg_parallel=True,
    run_perf_check=False,
),
```

评论区精华

关键讨论集中在 gemini-code-assist[bot] 指出的两个问题：

- parallelism_type 属性变更与 PR 描述矛盾：实际代码移除了 CFG_PARALLEL 返回逻辑，而描述称会返回 CFG_PARALLEL。reviewer 建议更新描述或将并行逻辑与调度器对齐。
- 新增的 CFG 组合逻辑与串行分支重复：建议抽取为 _cfg_combine 辅助方法（已被采纳实现）。 此外，mickqian 建议测试用例仅需 2 GPU 的 CFG parallel 配置即可，CloudRipple 补充说明 720p 用例仅用于路径验证。
- parallelism_type 返回值与 PR 描述矛盾 (correctness): 未明确回复，但代码保留为 REPLICATED，描述未更新。需验证调度器是否依赖此值。
- CFG 组合逻辑重复建议提取辅助方法 (design): 已采纳，实现了 _cfg_combine 方法。但串行分支尚未替换为调用该方法。
- 测试配置简化建议 (testing): 最终保留了 720p 单 GPU 和 360p CFG parallel 双 GPU 两个用例。

风险与影响

- 风险：
  1. 并行类型歧义：parallelism_type 始终返回 REPLICATED 可能影响调度器对阶段并行策略的判断，若调度器依赖 CFG_PARALLEL 进行资源分配，当前修改可能导致效率下降或错误。
  2. 代码重复隐患：虽已提取 _cfg_combine，但串行分支中仍保留着原始内联逻辑（未被替换为调用 _cfg_combine），存在后续维护不一致风险。
  3. 测试覆盖不足：新增测试仅验证基本路径是否通过，缺少对多 GPU 下精度和异常场景的校验。 - 影响：影响范围：MOVA 视频生成模型用户可直接受益于 CFG 并行带来的加速（benchmark 显示 e2e 从 211737ms 降至 110809ms）。影响程度：中等，仅限 diffusion 子模块，但引入的 _cfg_combine 辅助方法可复用于其他 diffusion 模型。团队需注意调度器与并行类型的语义一致性。 - 风险标记：并行类型语义矛盾，代码重复残留，测试覆盖不足

关联脉络

- PR #25683 [diffusion] feat: layerwise NVTX markers for Nsight Systems profiling: 同为 diffusion 模块的增强 PR，涉及性能分析工具，与本 PR 的多 GPU 优化配合可提升调试能力。

- PR #25964 [EPD] Cross-request batching for image/audio encoder: 涉及跨请求批处理的多 GPU 优化，与本 PR 的 CFG 并行设计理念类似，可参考其调度器交互模式。