PR #26925 完整报告

sgl-project/sglang

[multimodal_gen] Allow --dit-cpu-offload with --dit-layerwise-offload

合并时间：2026-06-01 23:17

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26925

执行摘要

- 一句话：允许 dit-cpu-offload 与 layerwise 共同启用
- 推荐动作：值得合并。修复了关键的启动失败问题，且测试覆盖完整。建议阅读 server_args.py 中 _adjust_layerwise_offload_components 和 _disable_non_dit_cpu_offload_for_layerwise_components 的变更，理解设计权衡。

功能与动机

当 --dit-layerwise-offload 启用时，ServerArgs 强制禁用 --dit-cpu-offload，导致低显存显卡（如 RTX 5090, 31.36 GiB 可用）加载大 DiT 模型（如 Qwen-Image-Edit-2509, 38.1 GiB）时 OOM。PR 中展示了详细的显存对比数据：两标志同时启用时启动成功，峰值 25499 MiB；禁用则 3 次 OOM。

实现拆解

1. 重命名并精简 _disable_cpu_offload_for_layerwise_components 方法：在 server_args.py 中，将原方法重命名为 _disable_non_dit_cpu_offload_for_layerwise_components，并移除其中强制关闭 dit_cpu_offload 的分支。现在该方法只关闭非 DiT 组件的 CPU offload（text_encoder、image_encoder、vae），因为只有这些组件会被 layerwise offload 接管。
2. 调整 _validate_offload 方法：移除以下两处强制关闭逻辑：一是当 layerwise offload 选中 DiT 组件时自动禁用 dit_cpu_offload 的警告和赋值；二是保留 __init__ 中 SGLANG_CACHE_DIT_ENABLED 的 ValueError 和 use_fsdp_inference 的 auto-disable（与 dit_cpu_offload 无关）。同时将局部变量 should_disable_dit_cpu_offload 重命名为 is_dit_layerwise_offload_selected 以反映其实际含义。
3. 更新 --dit-layerwise-offload 帮助文本：移除原有 "Cannot be used together with … dit_cpu_offload" 的描述，改为说明两者可组合使用，并指出组合后峰值显存最低。
4. 更新测试用例：
   - 将 test_layerwise_components_disable_matching_cpu_offloads 重命名为 test_layerwise_components_disable_matching_non_dit_cpu_offloads，并将断言 self.assertFalse(args.dit_cpu_offload) 改为 self.assertTrue(args.dit_cpu_offload)。
   - 新增 test_dit_layerwise_offload_preserves_dit_cpu_offload 回归测试，验证同时启用两个标志后值均保持为 True。

- 更新其他 5 个已有测试中的类似断言。

关键文件：

- python/sglang/multimodal_gen/runtime/server_args.py（模块 参数校验；类别 source；类型 core-logic；符号 _disable_cpu_offload_for_layerwise_components, _disable_non_dit_cpu_offload_for_layerwise_components, _validate_offload, add_cli_args）：核心变更文件：重命名方法并移除强制禁用 dit_cpu_offload 的逻辑，同时更新帮助文本。
- python/sglang/multimodal_gen/test/unit/test_server_args.py（模块 测试；类别 test；类型 test-coverage；符号 test_layerwise_components_disable_matching_cpu_offloads, test_layerwise_components_disable_matching_non_dit_cpu_offloads, test_dit_layerwise_offload_preserves_dit_cpu_offload）：配套测试更新：重命名测试方法并更新断言以验证 dit_cpu_offload 不被禁用，新增回归测试。

关键符号：_disable_non_dit_cpu_offload_for_layerwise_components, _validate_offload, add_cli_args, test_layerwise_components_disable_matching_non_dit_cpu_offloads, test_dit_layerwise_offload_preserves_dit_cpu_offload

关键源码片段

python/sglang/multimodal_gen/runtime/server_args.py

核心变更文件：重命名方法并移除强制禁用 dit_cpu_offload 的逻辑，同时更新帮助文本。

```
# python/sglang/multimodal_gen/runtime/server_args.py
# 关键变更：_disable_non_dit_cpu_offload_for_layerwise_components 不再修改 dit_cpu_offload

def _disable_non_dit_cpu_offload_for_layerwise_components(
    self, component_names: list[str]
) -> None:
    # non-DiT layerwise offload replaces the corresponding component-level CPU offload
    flag_names = cpu_offload_flags_for_layerwise_components(component_names)
    disabled_flag_names: list[str] = []
    # dit_cpu_offload 不再被强制设为 False，因为 dit_cpu_offload 与 layerwise offload
    # 互补：前者控制初始权重在主机内存，后者控制运行时逐层换入
    if (
        "text_encoder_cpu_offload" in flag_names
        and self.text_encoder_cpu_offload is not False
    ):
        self.text_encoder_cpu_offload = False
        disabled_flag_names.append("text_encoder_cpu_offload")
    if (
        "image_encoder_cpu_offload" in flag_names
        and self.image_encoder_cpu_offload is not False
    ):
        self.image_encoder_cpu_offload = False
        disabled_flag_names.append("image_encoder_cpu_offload")
    if "vae_cpu_offload" in flag_names and self.vae_cpu_offload is not False:
        self.vae_cpu_offload = False
```

```
            disabled_flag_names.append("vae_cpu_offload")
        explicit_disabled_flag_names = [
            flag_name
            for flag_name in disabled_flag_names
            if self.is_arg_explicitly_set(flag_name)
        ]
        if explicit_disabled_flag_names:
            logger.info(
                "Ignoring explicit CPU-offload flags because layerwise offload "
                "manages the same component weights: %s",
                ", ".join(
                    f"{flag_name}=False" for flag_name in explicit_disabled_flag_names
                ),
            )
```

python/sglang/multimodal_gen/test/unit/test_server_args.py

配套测试更新：重命名测试方法并更新断言以验证 dit_cpu_offload 不被禁用，新增回归测试。

```
# python/sglang/multimodal_gen/test/unit/test_server_args.py
# 新增回归测试：验证 --dit-cpu-offload 与 --dit-layerwise-offload 同时启用后均保持 True

def test_dit_layerwise_offload_preserves_dit_cpu_offload(self):
    """Combining --dit-cpu-offload with --dit-layerwise-offload must keep both on.

    dit_cpu_offload controls initial residency (host memory), while
    dit_layerwise_offload only swaps layers on/off device at inference.
    Force-disabling dit_cpu_offload here would push the full DiT to GPU at
    load time and OOM low-VRAM cards.
    """
    args = self._from_dict_with_task_type(
        ModelTaskType.T2I,
        memory_gb=32,
        kwargs={
            "dit_cpu_offload": True,
            "dit_layerwise_offload": True,
        },
    )

    self.assertTrue(args.dit_cpu_offload)
    self.assertTrue(args.dit_layerwise_offload)
    self.assertEqual(args.layerwise_offload_components, ["dit"])
```

评论区精华

Reviewer 评论（gemini-code-assist[bot]）：指出移除强制禁用代码后，变量 should_disable_dit_cpu_offload 名称已过时，建议重命名为 is_dit_layerwise_selected 或直接内联调用。该建议在后续提交（由 mickqian 完成）中被采纳。

- rename should_disable_dit_cpu_offload variable (style): 后续提交中（由 mickqian）重命名为 is_dit_layerwise_offload_selected，并内联使用。

风险与影响

- 风险：
  1. 回归风险：修改了配置验证逻辑，若其他代码路径依赖 dit_cpu_offload 被自动禁用，可能意外保留开启状态。但 PR 已通过单元测试覆盖主要组合，且低显存场景必须是同时启用才正常工作，回归概率较低。
  2. 性能影响：同时启用两者会增加额外的 H2D/D2H 开销（已在 GPU 逐层换入），但 loader 阶段不再 OOM，总体可用。PR body 中未提供推理阶段性能对比，建议关注。
  3. 兼容性风险：未引入新旧标志冲突，不影响仅使用单一标志的场景。 - 影响：用户：低显存 GPU（<40 GiB）用户可在不牺牲 DiT 层间卸载的前提下启用 CPU offload，避免加载时 OOM。系统：仅修改配置校验与帮助文本，运行时路径无变化。团队：修复了配置语义错误，提升了文档准确性。 - 风险标记：配置校验变更，显存敏感路径

关联脉络

- PR #26926 [diffusion] feat: improve cosmos3 serve API support: 同属 diffusion 模块，修改了 server_args.py 等文件，与当前 PR 有领域关联。