PR #21248 完整报告

sgl-project/sglang

[diffusion] Skip automatic Wan/MOVA DiT layerwise offload on high-end GPUs

合并时间：2026-03-25 18:45

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21248

执行摘要

此 PR 通过为高内存 GPU（≥130 GiB）自动禁用 Wan/MOVA DiT 层级卸载，优化了扩散模型的推理延迟。基于 H200 基准测试数据，避免了在高端 GPU 上引入性能回归，同时保持小内存 GPU 的原有平衡。

功能与动机

在单 H200 运行中，启用 dit_layerwise_offload=True 会导致延迟增加 60% 以上（如 Wan2.2-T2V-A14B 模型从 4.22s 到 6.77s）。为解决这一问题，PR 引入 130 GiB 内存阈值，当 CUDA 设备总内存≥130 GiB 时自动跳过卸载，以提升性能。动机源自 PR body 中的详细基准测试，显示卸载在高内存 GPU 上显著损害延迟。

实现拆解

改动集中在 python/sglang/multimodal_gen/runtime/server_args.py 文件：

- 添加常量：WAN_LAYERWISE_OFFLOAD_AUTO_DISABLE_MEM_GB = 130，作为内存阈值基准。
- 修改逻辑：在 _adjust_platform_specific 方法中，对于 Wan 或 MOVA 模型，当 dit_layerwise_offload 为 None 时，检查设备总内存；如果≥130 GiB，则设置 dit_layerwise_offload = False 并记录日志；否则自动启用卸载。
- 添加警告：在 _validate_offload 方法中增加彩色警告日志，提示卸载可能降低内存使用但增加延迟。

评论区精华

review 过程中没有技术讨论，两位 reviewer 直接批准，表明变更被认为合理且无争议。

风险与影响

风险：阈值设置依赖于有限基准测试，可能不适用于所有工作负载；内存检测准确性是关键依赖；缺少单元测试可能引入回归。 影响：高内存 GPU 用户自动获得更好延迟，但需确保阈值适应未来硬件；系统在低内存 GPU 上保持原有行为。

关联脉络

与 PR #21091（扩散模型性能 CI 测试）和 #21337（B200 性能绕过）相关，共同反映了团队在优化扩散模型性能、特别是针对高端 GPU 的持续努力。这些 PR 显示了性能调优和硬件适配的演进趋势。