PR #34952 完整报告

sgl-project/sglang

docs: define native diffusion model integration contract

合并时间：2026-08-15 23:37

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/34952

执行摘要

- 一句话：定义 diffusion 原生集成契约，明确 TP/SP、VAE 并行与 offload 要求
- 推荐动作：该 PR 值得作为 diffusion 新模型贡献的入门必读，也是 review diffusion 相关 PR 时的核对基准。值得关注的设计取舍包括：将分布式能力从“可选增强”提升为“完整支持的必要条件”；用共享 VAE 组件约束 tiled 与 spatial_shard 解码；以及明确 Diffusers 兼容后端与原生集成契约的边界。普通读者无需逐行精读，但建议维护者对照实际代码验证文档中引用的类名与行为。

功能与动机

PR body 明确指出需要澄清“完整原生 diffusion 模型支持”的边界，原文为：Clarify that complete native diffusion model support requires: encoder/DiT tensor parallelism and USP-based sequence parallelism; VAE tiled and spatial-shard decoding through shared VAE components; layerwise offload across loaded neural modules. The guide now includes the corresponding validation checks. This is a documentation-only change. 原有指南把多 GPU 支持写成“先单 GPU、之后再加 TP/SP”的可选步骤，容易让贡献者误以为单卡正确性即可合入，或允许用复制完整模型的方式绕过并行；本次将这三项能力提升为原生集成的必要条件并配套验证清单，减少后期返工与架构返修。

实现拆解

1. 重写第 5 节，把多 GPU 支持从可选升级为必选：将 ### 5. Denoising Module 改为 ### 5. Distributed and memory integration，删除“先单 GPU、之后再加 TP/SP”的表述，改为：encoder 与 DiT 必须原生支持 TP（原生并行投影与分片权重加载）与 SP（USPAttention），处理 mask、RoPE、padding 与输出聚合，禁止回退到复制完整模型，TP 与 SP 必须协同工作。
2. 补充 VAE 并行解码契约：要求子类化 ParallelTiledVAE 或复用满足同一契约的原生基类，通过 DecodingStage 与共享 decode group 支持 tiled 与 spatial_shard 解码，并指明复用 runtime/layers/parallel_conv.py 与 runtime/models/vaes/parallel/diffusers_spatial.py。这保证多 GPU 下 VAE 解码与 DiT 解码使用同一并行框架，避免各自为政。
3. 新增 layerwise offload 要求：所有加载的神经模块必须继承 LayerwiseOffloadableModuleMixin，并在 layer_names 中列出全部重复 block 路径；非 DiT 模块需设置 layerwise_offload_dit_group_enabled = False；并明确组件级 CPU offload 不能替代模块级分层 offload，避免内存优化与并行方案相互冲突。

4. 更新验证清单与勾选项：验证步骤新增 VAE 单独对比（含 tiled 与多 GPU spatial_shard）、encoder/DiT 的 TP、SP、TP x SP 组合以及 --layerwise-offload-components all 对照单 GPU 常驻基线的测试；原“Tested multi-GPU paths”复选框改为“完成上述分布式与内存集成检查”，使清单与第 5 节契约一一对应。
5. 明确 Diffusers 后端边界：文档把 Diffusers 后端定位为 compatibility-first，明确其不需要满足该原生集成契约，避免贡献者把兼容后端与原生实现混为一谈。

关键文件：

- docs/docs/sglang-diffusion/support_new_models.mdx（模块 集成文档；类别 docs；类型 documentation）：唯一变更文件；将 diffusion 新模型支持指南从单 GPU 视角提升为完整原生集成契约，并同步收紧验证清单，是后续所有 diffusion 模型合入的验收文档。

关键符号：未识别

评论区精华

该 PR 没有人工 review 评论，唯一的讨论是 mintlify[bot] 自动发布的 Mintlify 预览部署提示，未涉及技术设计：

> Preview deployment for your docs ... View Preview ... (mintlify[bot])

从提交历史可以看出作者自己做了两轮收紧：第一提交 docs: define native diffusion model integration contract 定义契约，第二提交 docs: tighten diffusion model integration guidance 进一步强化措辞（例如将 TP/SP 从“之后可加”改为“必须原生支持”，并补充 offload 细节），说明契约内容经过自我审查后更严谨。

- 暂无高价值评论线程

风险与影响

- 风险：风险主要来自文档与实现的一致性：契约中引用的 ParallelTiledVAE、LayerwiseOffloadableModuleMixin、USPAttention 等符号，若与当前代码命名或行为不一致，会直接误导贡献者；建议后续对照 runtime/models/vaes/parallel/、runtime/models/dits/ 等目录验证一遍。其次，把 TP/SP 与 offload 固化为硬性要求会提高新模型合入门槛，尤其是结构特殊、无法原生并行的架构；文档已明确 Diffusers 后端豁免，但后续仍需对 Diffusers 与原生两条路径的适用场景给出更清晰指引。本次为纯文档改动，无运行时影响，不存在回归、性能或安全风险。
- 影响：对用户无行为变化；对贡献者与维护者，diffusion 新模型合入验收标准从“单 GPU 正确 + 可选多卡”变为“TP/SP、VAE 并行、offload 三类检查齐全”，review 时可以本文档为核对清单。对系统不涉及代码、配置、依赖或部署，影响范围仅限文档目录 docs/docs/sglang-diffusion/。影响程度中等偏上：虽然代码零改动，但它重新定义了后续功能合入的契约边界，可能影响未来一段时间内 diffusion 相关 PR 的形态与返工量。
- 风险标记：文档契约或与实现脱节，新契约抬高贡献门槛

关联脉络

- PR #34891 fix(diffusion): scope attention backend fallback: 同为 diffusion 运行时层契约细化，涉及 attention 后端选择逻辑；文档引用的 qwen_image.py、wanvideo.py 等模型与其同属 diffusion 集成栈。
- PR #34825 [diffusion] Bound overlong weight lock filenames: 同为 diffusion 模块运行时的健壮性与契约性修正，与文档新增的 layerwise offload 模块约定相衔接。