

PR #30791 完整报告

sgl-project/sglang

[diffusion][docs] sync cookbook and log hygiene

合并时间: 2026-07-10 22:55

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30791>

执行摘要

- 一句话: 同步 diffusion 文档与降噪日志级别
- 推荐动作: 该 PR 为典型的文档与维护性变更, 可直接合并。适合所有关注 SGLang Diffusion 的用户阅读更新的文档, 了解最新功能参数。

功能与动机

PR body 明确指出需要同步文档以匹配当前 `performance_mode=speed`、compile warmup 和 breakable CUDA graph 行为; 同时通过合并 LingBot World 家族卡片和降级日志级别来提升文档清晰度和日志整洁性。

实现拆解

1. 文档入口更新 (docs_new/docs/sglang-diffusion/api/cli.mdx) : 在 CLI 参数列表中新增 `--enable-torch-compile`、`--offload-during-compile`、`--enable-breakable-cuda-graph`、`--bcg-text-buckets` 等参数的说明, 并更新 `--performance-mode` 中 `speed` 模式的行为描述。
2. 部署手册同步 (docs_new/docs/sglang-diffusion/deployment_cookbook.mdx) : 补充 `speed` 模式下 `torch compile` 默认启用、编译期间组件卸载、以及 `breakable CUDA graph` 手动启用的说明。
3. 概览页重构 (docs_new/cookbook/diffusion/intro.mdx) : 将 LingBot World 和 LingBot World 2.0 的两张卡片合并为一张, 并将链接指向较新的 2.0 食谱, 同时添加说明引导用户使用侧边栏浏览家族变体。
4. README 刷新 (python/sglang/multimodal_gen/README.md) : 更新支持的模型列表、加速特性描述及多平台列表, 新增 Intel XPU、Ascend NPU 等平台。
5. 日志降噪 (python/sglang/multimodal_gen/runtime/pipelines_core/stages/model_specific_stages/lingbot_world/lingbot_world_causal_denoising.py) : 将 `_log_lingbot_kv_window` 方法中的 `logger.info` 改为 `logger.debug`, 降低交互式 KV 窗口日志频率。

关键文件:

- python/sglang/multimodal_gen/runtime/pipelines_core/stages/model_specific_stages/lingbot_world/lingbot_world_causal_denoising.py (模块 日志管理; 类别 source; 类型 data-contract; 符号 `_log_lingbot_kv_window`) : 唯一源码修改: 将 `logger.info` 改为

logger.debug, 降低日志级别。

- docs_new/docs/sglang-diffusion/api/cli.mdx (模块 文档; 类别 other; 类型 entrypoint) : 主要文档更新, 新增多个 CLI 参数说明。
- docs_new/docs/sglang-diffusion/deployment_cookbook.mdx (模块 文档; 类别 other; 类型 core-logic) : 更新部署手册, 补充 speed 模式与 torch.compile 等行为。
- docs_new/cookbook/diffusion/intro.mdx (模块 文档; 类别 other; 类型 core-logic) : 合并 LingBot World 卡片, 简化导航。
- python/sglang/multimodal_gen/README.md (模块 文档; 类别 docs; 类型 documentation) : 刷新支持的模型与平台列表。

关键符号: _log_lingbot_kv_window

关键源码片段

[python/sglang/multimodal_gen/runtime/pipelines_core/stages/model_specific_stages/lingbot_world/lingbot_world_causal_denoising.py](#)

唯一源码修改: 将 logger.info 改为 logger.debug, 降低日志级别。

```
# lingbot_world_causal_denoising.py
# 降低日志级别以减少生产噪音
logger.debug( # 原来为 logger.info
    "LingBot interactive KV window: session_id=%s request_id=%s "
    "chunk_idx=%s mode=%s window_frames=%s sample_frames=%s "
    "cache_frames=%s sink_frames=%s current_frames=%s sample_tokens=%s "
    "cache_tokens=%s still_chunks=%s",
    getattr(batch, "realtime_session_id", None),
    getattr(batch, "request_id", None),
    getattr(batch, "block_idx", None),
    mode,
    window_frames,
    sample_frames,
    int(self.sliding_window_num_frames),
    int(self.sink_size),
    int(self.num_frames_per_block),
    sample_tokens,
    int(self.sliding_window_num_frames) * int(self.num_token_per_frame),
    still_chunks,
)
```

评论区精华

该 PR 无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅涉及文档和日志级别的调整。日志级别从 info 降为 debug 不会影响功能，但可能使调试时信息不可见，需确保日志系统配置可捕获 debug 级别。
- 影响：对用户：文档更新提高配置可见性，日志降级减少噪音。对系统：无功能影响。对团队：维护成本低。影响程度主要为文档和用户体验。
- 风险标记：日志级别修改可能影响调试

关联脉络

- PR #30584 Fix diffusion BCG lifetime and add Z-Image-Turbo CI: 该 PR 涉及 breakable CUDA graph 的修复与 CI 测试，与本次文档中新增的 breakable CUDA graph 参数说明相关。
- PR #30447 [diffusion] Reorganize runtime utility modules: 该 PR 对 diffusion 运行时模块进行了重组，可能影响文档中引用的组件。