

PR #24677 完整报告

sgl-project/sglang

[AMD] Intro SGLANG_DIFFUSION_AITER_FP8_ATTN

合并时间: 2026-05-08 16:31

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/24677>

执行摘要

- 一句话: 重命名 FP8 注意力环境变量为 diffusion 专属
- 推荐动作: 建议精读以了解环境变量命名规范。当前 PR 未提供向后兼容, 团队应尽快跟进更新相关部署配置, 或在后续 PR 中添加 fallback 支持 (如 new_var 未设置时读取 old_var)。

功能与动机

原环境变量名 `SGLANG_AITER_FP8_ATTN` 含义过于宽泛, 可能被误用于非 diffusion 场景。更名为 `SGLANG_DIFFUSION_AITER_FP8_ATTN` 以明确该开关仅控制 diffusion 模型的 Aiter FP8 注意力。

实现拆解

1. 环境变量重命名: 在 `python/sglang/multimodal_gen/runtime/layers/attention/backends/aiter.py` 中, 将 `_use_fp8_attn` 的获取键从 `"SGLANG_AITER_FP8_ATTN"` 改为 `"SGLANG_DIFFUSION_AITER_FP8_ATTN"`, 默认值仍为 `"0"`。
2. 日志更新: 同一文件中, `logger.info` 的消息内容同步更新为新变量名。
3. 文档注释更新: `forward` 方法 `docstring` 中提及的环境变量名更新为新名称。
4. 测试同步: 在 `test/registered/amd/test_wan22_fp8_mla.py` 中, `FP8_ENV` 字典的键从 `"SGLANG_AITER_FP8_ATTN"` 改为 `"SGLANG_DIFFUSION_AITER_FP8_ATTN"`, 保持测试与源码一致。

关键文件:

- `python/sglang/multimodal_gen/runtime/layers/attention/backends/aiter.py` (模块 注意力后端; 类别 source; 类型 core-logic): 核心源码文件, 环境变量名在此定义和引用, 直接控制 FP8 注意力是否启用。
- `test/registered/amd/test_wan22_fp8_mla.py` (模块 AMD 测试; 类别 test; 类型 test-coverage): 测试文件, 同步更新环境变量以保持测试有效性。

关键符号: 未识别

关键源码片段

`python/sglang/multimodal_gen/runtime/layers/attention/backends/aiter.py`

核心源码文件，环境变量名在此定义和引用，直接控制 FP8 注意力是否启用。

```
# python/sglang/multimodal_gen/runtime/layers/attention/backends/aiter.py
# 环境变量键重命名为 diffusion 专属，避免歧义
_use_fp8_attn = os.environ.get("SGLANG_DIFFUSION_AITER_FP8_ATTN", "0") == "1"
_fp8_dtype = torch.float8_e4m3fn

if _use_fp8_attn:
    logger.info("DiT FP8 attention enabled via SGLANG_DIFFUSION_AITER_FP8_ATTN=1")
```

评论区精华

gemini-code-assist[bot] 提出兼容性问题：直接重命名会破坏依赖旧变量名的已有工作流和 CI 配置，建议同时支持两个变量，以新变量优先。但作者未采纳，未提供 fallback。

- 向后兼容性：是否应同时支持旧环境变量 (design)：作者未采纳，仅重命名为新变量，未提供 fallback。

风险与影响

- 风险：破坏性变更风险：任何依赖原环境变量 SGLANG_AITER_FP8_ATTN 的部署脚本、CI 配置或用户启动命令将不再生效，FP8 注意力会静默回退到 BF16（因新变量未设置），可能导致性能下降但无错误提示，排查困难。该风险程度中等。
- 影响：
 - 用户：使用 AMD GPU 并启用 FP8 MLA 注意力的 diffusion 用户需要更新环境变量设置。
 - 系统：仅影响 diffusion 模型推理路径中的 Aiter 注意力后端，无隐式影响。
 - 团队：需要同步更新 CI、Dockerfile 和文档中对旧变量名的引用。
 - 风险标记：破坏性变更（环境变量更名无 fallback），静默回退风险（未设置新变量时无警告）

关联脉络

- PR #20319 [AMD] Support fp8 MLA for diffusion model: 引入了 SGLANG_AITER_FP8_ATTN 环境变量，本 PR 将其重命名为更精确的名称。
- PR #23955 [AMD] Add AMD FP8 MLA attention test for Wan2.2-T2V-A14B: 相关测试文件，本 PR 同步更新了其中的环境变量名。