

PR #36019 完整报告

sgl-project/sglang

[diffusion] Honor XDG cache for model overlays

合并时间: 2026-08-24 14:11

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36019>

执行摘要

- 一句话: 模型 overlay 缓存路径改为优先服从 XDG 缓存目录
- 推荐动作: 值得快速浏览, 变更小而聚焦。重点理解 `get_diffusion_cache_root()` 为何改为从 `envs` 取值, 以及为何默认路径改为 `XDG_CACHE_HOME`。该改动体现了对 Linux 缓存规范的遵循。

功能与动机

在 H200 批量扫描中, LTX-2.3 在 `/root/.cache/sgl_diffusion` 下留下了约 146GB 和 31 个权重文件, 尽管 `XDG_CACHE_HOME` 指向了任务缓存。这导致 overlay 检查点 (如 LTX-2.3 和 JoyEcho) 将权重物化在任务隔离缓存之外, 造成资源泄漏和潜在的缓存污染。PR 旨在通过遵循 XDG 缓存规范来隔离缓存, 防止此类泄漏。

实现拆解

1. 修改缓存根目录解析函数: 在 `python/sglang/multimodal_gen/runtime/utils/model_overlay.py` 中, 将 `get_diffusion_cache_root()` 的实现从直接读取环境变量改为引用 `envs.SGLANG_DIFFUSION_CACHE_ROOT`, 该变量封装了 `XDG_CACHE_HOME` 的默认回退逻辑。
2. 引入 `envs` 模块依赖: 在文件头部新增 `from sglang.multimodal_gen import envs` 导入, 确保环境变量解析逻辑与模块其他部分保持一致。
3. 新增单元测试: 在 `python/sglang/multimodal_gen/test/unit/test_model_overlay.py` 中新增两个测试用例, 分别验证默认回退到 XDG 缓存目录和显式设置覆盖优先级的场景。

通过这些改动, 缓存路径解析逻辑集中到 `envs` 中, 后续修改环境变量行为只需调整一处。

关键文件:

- `python/sglang/multimodal_gen/runtime/utils/model_overlay.py` (模块 缓存路径; 类别 source; 类型 data-contract; 符号 `get_diffusion_cache_root`): 核心源码修改, 将缓存根目录解析改为复用 `envs` 模块, 使其遵循 `XDG_CACHE_HOME`
- `python/sglang/multimodal_gen/test/unit/test_model_overlay.py` (模块 测试; 类别 test; 类型 test-coverage; 符号 `test_uses_xdg_cache_home_by_default`, `test_explicit_diffusion_cache_root_takes_precedence`): 新增测试, 覆盖默认 XDG 回退和显式设置优先级, 保障行为正确性

关键符号: `get_diffusion_cache_root`

关键源码片段

`python/sglang/multimodal_gen/runtime/utils/model_overlay.py`

核心源码修改, 将缓存根目录解析改为复用 `envs` 模块, 使其遵循 `XDG_CACHE_HOME`

```
# python/sglang/multimodal_gen/runtime/utils/model_overlay.py
# 通过复用 envs 模块, 统一缓存根目录解析逻辑, 默认回退到 XDG_CACHE_HOME/sgl_diffusion
from sglang.multimodal_gen import envs
```

```
def get_diffusion_cache_root() -> str:
    # 直接返回 envs 中解析好的值:
    # - 显式设置 SGLANG_DIFFUSION_CACHE_ROOT 时优先
    # - 否则回退到 XDG_CACHE_HOME/sgl_diffusion (默认 XDG_CACHE_HOME 为 ~/.cache)
    return envs.SGLANG_DIFFUSION_CACHE_ROOT
```

`python/sglang/multimodal_gen/test/unit/test_model_overlay.py`

新增测试, 覆盖默认 XDG 回退和显式设置优先级, 保障行为正确性

```
# python/sglang/multimodal_gen/test/unit/test_model_overlay.py
from sglang.multimodal_gen.runtime.utils.model_overlay import get_diffusion_cache_root

def test_uses_xdg_cache_home_by_default(monkeypatch):
    # 仅设置 XDG_CACHE_HOME, 未显式设置 diffusion 缓存根时, 应回退到 XDG 缓存目录下
    monkeypatch.setenv("XDG_CACHE_HOME", "/tmp/sglang-xdg-cache")
    monkeypatch.delenv("SGLANG_DIFFUSION_CACHE_ROOT", raising=False)

    assert get_diffusion_cache_root() == "/tmp/sglang-xdg-cache/sgl_diffusion"

def test_explicit_diffusion_cache_root_takes_precedence(monkeypatch):
    # 显式设置 diffusion 缓存根时, 应覆盖 XDG_CACHE_HOME
    monkeypatch.setenv("SGLANG_DIFFUSION_CACHE_ROOT", "/tmp/sglang-diffusion-cache")
    monkeypatch.setenv("XDG_CACHE_HOME", "/tmp/sglang-xdg-cache")

    assert get_diffusion_cache_root() == "/tmp/sglang-diffusion-cache"
```

评论区精华

该 PR 没有评论或审核讨论。唯一的评论是作者 BBuf 触发的 CI 重新运行指令。

- 暂无高价值评论线程

风险与影响

- 风险：变更影响所有 diffusion 模型 overlay 的缓存路径解析。主要风险是： 1) 对已有用户，如果之前依赖默认路径 ~/.cache/sgl_diffusion，且未设置 XDG_CACHE_HOME，则行为不变（因为 XDG 默认值就是 ~/.cache）； 2) 如果设置了 XDG_CACHE_HOME，缓存位置会变化，可能导致重新下载权重，但这是预期行为； 3) 由于逻辑移入 envs 模块，需要确保 envs 模块的导入不会引起循环依赖，目前看来依赖关系清晰。
- 影响：影响所有使用 diffusion 模型 overlay 的用户，尤其是设置 XDG_CACHE_HOME 的环境。修复后缓存位置更规范，避免了权重泄漏到隔离缓存之外，减少了磁盘占用和潜在冲突。对系统而言，缓存路径解析逻辑统一到 envs，提高了可维护性。对团队而言，新增的测试确保了行为回归。影响程度中等，属于行为修正。
- 风险标记：缓存路径行为变化，可能需重新下载权重，依赖 envs 模块实现细节

关联脉络

- PR #36062 [diffusion] cache LoRA-merged weights in files the page cache can hold: 同属 diffusion 模块的缓存管理改进，关注缓存位置与性能