

PR #28443 完整报告

sgl-project/sglang

[XPU] fix(deps): upgrade diffusers to fix fresh installs

合并时间: 2026-06-18 12:44

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28443>

执行摘要

- 一句话: 升级 XPU 平台 diffusers 依赖版本
- 推荐动作:
 1. 精简有效: PR 直接命中问题根因而改动最小, review 后及时移除了不需要的 `apache-tvm-ffi` 依赖, 可快速合入。
 2. 值得精读: 对于其他平台维护者, 可参考此 PR 的逻辑, 将 diffusers 升级同步到对应的 `pyproject` 文件中。
 3. 设计决策: 保持了 XPU 平台的独立配置, 未盲目同步到所有平台, 体现了谨慎的变更策略。

功能与动机

XPU 平台新安装环境运行 diffusion 模型时, 由于 `diffusers==0.36.0` 中 `torchao_quantizer.py` 从 `torchao.dtypes.floatx.float8_layout` 导入 `Float8AQTTensorImpl` (CUDA 专属路径, XPU torchao 不存在), 且异常处理中引用了未定义的 `logger` 变量, 导致 `ModuleNotFoundError`。此问题在 `diffusers>=0.38.0` 中已修复。

实现拆解

1. 定位依赖问题: 在 `python/pyproject_xpu.toml` 的 `[project.optional-dependencies]` 下 `diffusion` 组中, 将 `diffusers==0.36.0` 改为 `diffusers==0.38.0`。此举仅更改版本号, 不影响其他依赖。
2. 移除不必要依赖: 最初提交尝试添加 `apache-tvm-ffi==0.1.9`, 但经 review 指出去重后的 PR#28426 已移除相关代码, 现无需此依赖, 第二笔提交将其还原。
3. 验证修复: 作者提供命令行测试, 使用 `diffusers==0.38.0` 运行 `sglang generate` (模型 `Tongyi-MAI/Z-Image-Turbo`) 确认正常。

关键文件:

- `python/pyproject_xpu.toml` (模块 依赖配置; 类别 `config`; 类型 `configuration`): 唯一变更文件, 将 `diffusers` 版本从 `0.36.0` 升级到 `0.38.0`, 修复 XPU 平台 diffusion 模型加载失败问题。

关键符号: 未识别

关键源码片段

python/pyproject_xpu.toml

唯一变更文件，将 diffusers 版本从 0.36.0 升级到 0.38.0，修复 XPU 平台 diffusion 模型加载失败问题。

```
# python/pyproject_xpu.toml
[project.optional-dependencies]
diffusion = [
    "PyYAML==6.0.1",
    "cloudpickle==3.1.2",
    "diffusers==0.38.0", # 升级从 0.36.0 到 0.38.0，修复 XPU torchao 兼容性和 logger 未定义 bug
    "imageio==2.36.0",
    "imageio-ffmpeg==0.5.1",
    "moviepy>=2.0.0",
    "opencv-python==4.10.0.84",
    "remote-pdb==2.1.0",
    "st_attn==0.0.7 ; platform_machine != 'aarch64' and platform_machine != 'arm64'",
    "runai_model_streamer>=0.15.5",
    "cache-dit==1.3.0",
    "addict==2.4.0",
    "av==16.1.0",
    "scikit-image==0.25.2",
    "trimesh>=4.0.0",
    "xatlas",
]
```

评论区精华

1. apache-tvm-ffi 依赖是否必要 (polisettyvarma、ck-intel)

- polisettyvarma 指出最新代码中可能不再需要 apache-tvm-ffi，建议核验。
- ck-intel 确认该依赖已被 PR#28426 回滚，因此无需添加。最终从 PR 中移除。

2. 是否同步更新其他平台配置文件 (gemini-code-assist[bot])

- 机器人建议将 diffusers 升级同步到 pyproject_cpu.toml、pyproject_npu.toml 等文件，以保持一致性。
- 但此 PR 仅针对 XPU 平台，且其他平台可能尚未遇到或已单独处理，因此未采纳。
- apache-tvm-ffi 依赖必要性 (correctness): 移除 apache-tvm-ffi 依赖，仅保留 diffusers 版本升级。
- 同步更新其他平台配置文件 (other): 未采纳，仅针对 XPU 平台修复。

风险与影响

• 风险:

1. 回归风险: diffusers 从 0.36.0 升级到 0.38.0 属于 minor 版本跳跃，但用户测试验证了基本场景，风险较低。主要改动在依赖声明，无代码逻辑变更。

2. 兼容性风险: diffusers 0.38.0 可能与 SGLang 中其他扩散模型功能存在兼容问题, 但作者仅测试了一个模型, 覆盖范围有限。

3. 其他平台: 仅修复 XPU 平台, CPU/NPU 等若存在相同问题需单独处理。

- 影响:

- 用户影响: XPU 平台用户不再因 diffusers 版本问题导致 diffusion 模型启动失败, 新安装环境可正常使用。

- 系统影响: 仅修改 pyproject_xpu.toml 一行, 无运行时性能或稳定性影响。

- 团队影响: 维护成本极低, 但未同步更新其他平台文件可能导致未来出现类似的 issue。

- 风险标记: 缺少跨平台同步

关联脉络

- PR #28426 Revert import of tvn_ffl in dsv4: 移除了对 apache-tvm-ffi 的 import, 使得本 PR 中无需添加该依赖。