PR #25451 完整报告

sgl-project/sglang

Upgrade transformers to 5.8.1

合并时间：2026-05-19 22:20

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25451

执行摘要

- 一句话：统一升级 transformers 到 5.8.1
- 推荐动作：建议合并此 PR，但密切关注 CI 测试结果，特别是模型加载与 tokenizer 相关测试；若出现失败，应优先排查 transformers 5.8.1 的 breaking changes。后续可考虑补充针对 transformers 版本的集成测试。

功能与动机

PR body 明确为 'upgrade transformers pins to the latest PyPI release, 5.8.1' 并 'keep SGLang package variants and the AMD wheel copy aligned'，目的是跟随上游最新稳定版本，获取 bug 修复与性能改进。

实现拆解

1. 遍历 6 个 pyproject.toml 文件，找到 transformers 依赖声明。
2. 将版本值统一改为 ==5.8.1，其中主 python/pyproject.toml、CPU、NPU、Other、XPU 从 5.6.0 升级，AMD 从 4.57.1 升级。
3. 未改动任何源代码、测试或配置结构，仅版本号变更。

关键文件：

- python/pyproject.toml（模块 主包配置；类别 config；类型 configuration）：主包依赖定义，影响所有 CUDA 用户；版本从 5.6.0 升级到 5.8.1。
- 3rdparty/amd/wheel/sglang/pyproject.toml（模块 AMD 配置；类别 config；类型 configuration）：AMD 专用依赖定义，版本从 4.57.1 直接跳升至 5.8.1，跨度最大。
- python/pyproject_cpu.toml（模块 CPU 配置；类别 config；类型 configuration）：CPU 变体依赖，同步升级保证一致性。
- python/pyproject_npu.toml（模块 NPU 配置；类别 config；类型 configuration）：NPU 变体依赖，同步升级保证一致性。
- python/pyproject_other.toml（模块 其他配置；类别 config；类型 configuration）：其他平台（如无特殊标识）的依赖，保持对齐。
- python/pyproject_xpu.toml（模块 XPU 配置；类别 config；类型 configuration）：XPU 变体依赖，同步升级保证一致性。

关键符号：未识别

关键源码片段

python/pyproject.toml

主包依赖定义，影响所有 CUDA 用户；版本从 5.6.0 升级到 5.8.1。

```
# python/pyproject.toml (dependencies section, partial)
# 关键变更：transformers==5.6.0 → transformers==5.8.1
# 保持与 torch 2.11.0、torchao 0.17.0 等兼容

dependencies = [
    # ... 省略其他依赖 ...
    "mistral_common>=1.11.0",
    "transformers==5.8.1", # 从 5.6.0 升级至最新 PyPI 版
    "uvicorn",
    "uvloop",
    "watchfiles",
    "xgrammar==0.2.0",
    "smg-grpc-servicer>=0.5.0",
    "kernels",
]
```

3rdparty/amd/wheel/sglang/pyproject.toml

AMD 专用依赖定义，版本从 4.57.1 直接跳升至 5.8.1，跨度最大。

```
# 3rdparty/amd/wheel/sglang/pyproject.toml (runtime_common section)
# 关键变更：transformers==4.57.1 → transformers==5.8.1
# AMD ROCm 特定依赖
runtime_common = [
    # ... 省略 ...
    "timm==1.0.16",
    "torchao==0.9.0",
    "tqdm",
    "transformers==5.8.1", # 从 4.57.1 升级，注意兼容性
    "uvicorn",
    "uvloop",
    "xgrammar==0.2.0",
]
```

评论区精华

无实质性讨论；仅有 1 位 reviewer (ispobock) 直接 APPROVED，未提出任何问题或评论。

- 暂无高价值评论线程

风险与影响

- 风险：主要风险来自 transformers 5.6.0 → 5.8.1 的跳级升级，可能引入 API 废弃、行为变化或与现有模型加载代码的不兼容。由于未同步调整任何源码或测试，若 CI 测试覆盖不全，可能遗漏回归。AMD 变体从 4.57.1 跳升至 5.8.1 跨度更大，需额外关注。

- 影响：影响所有使用 SGLang 的用户（包括 CPU、NPU、AMD、XPU 等平台），因为 transformers 是模型加载、tokenizer 等核心依赖。预期补丁兼容，但建议用户验证关键模型功能。对开发团队而言，此 PR 统一了版本，降低了维护负担。
- 风险标记：核心依赖升级，无代码适配，跳过测试覆盖

关联脉络

- PR #23922 transformers v5 adapt HFRunner: 该 PR 为 transformers v5 做了代码适配，本次升级进一步推进版本，有前后演进关系。