PR #28483 完整报告

sgl-project/sglang

[Fix] DeepSeek-OCR-2 bench_serving: fix processor loading

合并时间：2026-06-18 06:29

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28483

执行摘要

- 一句话：修复 DeepSeek-OCR-2 bench_serving 处理器加载
- 推荐动作：该 PR 改动清晰、风险低，值得合并。但建议后续补充对应 DeepSeek-OCR-2 的 benchmark 测试用例，防止回归。团队可关注 hf_transformers_utils.get_processor 对更多模型的处理逻辑，避免类似问题。

功能与动机

Bug prevented bench_serving from working correctly with DeepSeek-OCR-2. The original get_processor in benchmark/utils.py called AutoProcessor.from_pretrained directly, bypassing the model-specific processor detection in hf_transformers_utils.get_processor. This caused the wrong processor class (DeepseekVLV2Processor) to be loaded for OCR-2, resulting in None being passed into string operations during token counting.

实现拆解

1. 统一 import 路径：在 python/sglang/benchmark/utils.py 的 get_processor 函数中，将原有的按后缀名（.json/.model）分支导入 hf_transformers_utils.get_processor 的方式，改为在最顶部统一导入并重命名为 _srt_get_processor（第 75-77 行），消除了重复导入。
2. 重构路径判断逻辑：将原有的 if-elif 结构（先判断是否为 .json/.model 文件来决定是否走 hf_transformers_utils，否则下载模型再调用 AutoProcessor.from_pretrained）简化为一个统一的 if 条件（第 79-82 行）：如果路径不以 .json/.model 结尾且本地不存在，则下载模型。不再需要针对 .json/.model 做特殊处理分支，因为 hf_transformers_utils.get_processor 内部已经包含了相关逻辑。
3. 替换最终调用：将最后的 return AutoProcessor.from_pretrained(...)（第 85-87 行）替换为 return _srt_get_processor(...)（第 83 行），使得所有情况都通过 hf_transformers_utils.get_processor 获取处理器，该函数会检查内部维护的 _CUSTOMIZED_MM_PROCESSOR 字典，为 DeepSeek-OCR-2 等模型返回正确的处理器类，对其他模型则回退到 AutoProcessor.from_pretrained。
4. 无测试变更：本次仅修改源码，未添加对应单元测试。

关键文件：

- python/sglang/benchmark/utils.py（模块 基准测试；类别 source；类型 dependency-wiring；符号 get_processor）：唯一的变更文件，核心修复了 get_processor 函数，将处理器获取逻辑统一到 hf_transformers_utils.get_processor，修复了 DeepSeek-OCR-2 模型在 benchmark 中崩溃的问题。

关键符号：get_processor

关键源码片段

python/sglang/benchmark/utils.py

唯一的变更文件，核心修复了 get_processor 函数，将处理器获取逻辑统一到 hf_transformers_utils.get_processor，修复了 DeepSeek-OCR-2 模型在 benchmark 中崩溃的问题。

```
def get_processor(
    pretrained_model_name_or_path: str,
) -> AutoProcessor:
    assert (
        pretrained_model_name_or_path is not None
        and pretrained_model_name_or_path != ""
    )

    # 统一导入 hf_transformers_utils 中的 get_processor
    # 该函数内部维护了自定义处理器映射（如 DeepSeek-OCR-2），
    # 对于标准模型则回退到 AutoProcessor.from_pretrained
    from sglang.srt.utils.hf_transformers_utils import (
        get_processor as _srt_get_processor,
    )

    # 只有当路径不是 .json/.model 文件且本地不存在时，才调用 get_model 下载
    if not pretrained_model_name_or_path.endswith(
        (".json", ".model")
    ) and not os.path.exists(pretrained_model_name_or_path):
        pretrained_model_name_or_path = get_model(pretrained_model_name_or_path)

    # 始终通过 _srt_get_processor 获取处理器，确保模型特定的分支被正确触发
    return _srt_get_processor(pretrained_model_name_or_path, trust_remote_code=True)
```

评论区精华

Review 中主要讨论集中在 CI 环境问题。mingfeima 指出 XPU CI 失败，但怀疑是 CI 环境本身问题而非代码变更导致。其余无关于代码实现的实质性讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。该变更将 get_processor 统一导向 hf_transformers_utils.get_processor，该函数已在 SRT 服务端广泛使用且设计上兼容所有模型（对未知模型回退到

AutoProcessor.from_pretrained）。唯一风险是该函数可能依赖额外的 HF 库版本特性，但已在 benchmark/utils.py 中预先导入，不会影响已有逻辑。

- 影响：直接影响：修复了 DeepSeek-OCR-2 模型在 bench_serving 中因处理器加载错误而崩溃的问题。间接影响：所有其他模型（如 LLaVA、Qwen-VL、InternVL、Phi-3-Vision 等）的 benchmark 行为完全不变，因为 hf_transformers_utils.get_processor 对非定制模型会回退到原有 AutoProcessor.from_pretrained 逻辑。影响范围限定在 benchmark 工具内。
- 风险标记：缺少测试覆盖

关联脉络

- 暂无明显关联 PR