

PR #29789 完整报告

sgl-project/sglang

chore: clean diffusion dead code

合并时间: 2026-07-01 15:42

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29789>

执行摘要

- 一句话: 清理 diffusion 模块的未使用导入和死代码
- 推荐动作: 作为常规清理 PR, 不必强制精读; 但 vmoba/__init__.py 中的显式 re-export 模式值得在其他模块推广, 可提升 IDE 跳转和静态分析准确性。

功能与动机

从 PR Body 和提交说明看, 主要目标是清除 diffusion 子目录中累积的死代码: 未使用的导入 / 局部变量、被注释掉的旧 attention 实现、空 f-string 以及低价值的 TODO/smoke 表述, 同时保持导出接口的清晰性, 提高代码可维护性。

实现拆解

1. 移除未使用的导入: 通过 `ruff check --select F401` 识别并删除 `qwen2_5vl.py` 中 `Callable` 和 `eager_attention_forward` 等不再需要的 import。
2. 删除注释掉的 attention 代码: 在 `qwen2_5vl.py` 的 `forward` 方法中, 移除被 `#` 注释的备用 attention 实现以及 `attention_interface` 变量声明。
3. 清理 VAE 中的 deprecation 消息: 在 `autoencoder.py` 和 `autoencoder_kl_flux2.py` 的 `tilde_encode` 方法中, 删除未实际调用的 `deprecation_message` 字符串和 `# deprecate(...)` 注释。
4. 显式化 re-export 的语义: 将 `vmoba/__init__.py` 和 `pipelines_core/__init__.py` 中的 `from .x import y` 改写为 `from .x import y as y` 并补充 `__all__`, 以便静态分析工具正确识别符号来源。
5. 清除低信号 TODO 和空 f-string: 在 `qwen_image.py`、`glm_image.py`、`joy_image.py` 等文件中移除注释掉的代码行和空格式字符串。

关键文件:

- `python/sglang/multimodal_gen/runtime/models/encoders/qwen2_5vl.py` (模块 扩散模型; 类别 source; 类型 import-cleanup): 移除了未使用的 `Callable` 导入、`eager_attention_forward` 导入以及被注释掉的 attention 备用路径, 是清理力度最大的文件。
- `python/sglang/multimodal_gen/runtime/models/vaes/autoencoder.py` (模块 扩散模型; 类别 source; 类型 code-cleanup): 删除了 `tilde_encode` 中从未执行的 `deprecation_message` 字符串和注释掉的 `deprecate` 调用, 减少代码噪音。

- python/sglang/multimodal_gen/runtime/models/vaes/autoencoder_kl_flux2.py (模块 扩散模型; 类别 source; 类型 code-cleanup) : 与 autoencoder.py 类似, 移除了 deprecation 消息和未使用的 force_upcast 变量。
- python/sglang/multimodal_gen/csrc/attn/vmoba_attn/vmoba/__init__.py (模块 扩散模型; 类别 source; 类型 dependency-wiring) : 将隐式 re-export 改为显式 as 形式并增加 __all__, 改善模块接口的可维护性。
- python/sglang/multimodal_gen/runtime/pipelines_core/__init__.py (模块 扩散模型; 类别 source; 类型 dependency-wiring) : 与 vmoba 类似, 将两个导入改为显式 as 形式, 保持一致性。

关键符号: forward, tiled_encode

关键源码片段

python/sglang/multimodal_gen/runtime/models/encoders/qwen2_5vl.py

移除了未使用的 `Callable` 导入、`eager_attention_forward` 导入以及被注释掉的 `attention` 备用路径, 是清理力度最大的文件。

```
# qwen2_5vl.py - 清理后的 forward 方法 (关键改动)
def forward(
    self,
    hidden_states: torch.Tensor,
    position_embeddings: Optional[tuple[torch.Tensor, torch.Tensor]] = None,
    past_key_values: Optional[...] = None,
    cache_position: Optional[torch.LongTensor] = None,
    **kwargs: Unpack[FlashAttentionKwargs],
) -> tuple[torch.Tensor, Optional[torch.Tensor], Optional[tuple[torch.Tensor]]]:
    # ... (省略前处理)

    # 删除了 `attention_interface: Callable = eager_attention_forward` 及其下方的注释块
    query_states = query_states.transpose(1, 2)
    key_states = key_states.transpose(1, 2)
    value_states = value_states.transpose(1, 2)
    attn_output = self.attn(query_states, key_states, value_states)

    attn_output = attn_output.reshape(bsz, q_len, -1).contiguous()
    attn_output = _linear_output(self.o_proj, attn_output)
    return attn_output
```

python/sglang/multimodal_gen/runtime/models/vaes/autoencoder.py

删除了 `tiled_encode` 中从未执行的 `deprecation_message` 字符串和注释掉的 `deprecate` 调用, 减少代码噪音。

```
# autoencoder.py - 清理后的 tiled_encode 方法 (开头部分)
def tiled_encode(
    self, x: torch.Tensor, return_dict: bool = True
) -> AutoencoderKLOutput:
    r"""Encode a batch of images using a tiled encoder.
```

```
...(文档字符串不变)
"""
# 已删除: deprecation_message 字符串和 # deprecate(...) 注释
overlap_size = int(self.tile_sample_min_size * (1 - self.tile_overlap_factor))
blend_extent = int(self.tile_latent_min_size * self.tile_overlap_factor)
row_limit = self.tile_latent_min_size - blend_extent
# ... 后续代码不变
```

python/sglang/multimodal_gen/csrc/attn/vmoba_attn/vmoba/__init__.py

将隐式 re-export 改为显式 as 形式并增加 `__all__`，改善模块接口的可维护性。

```
# vmoba/__init__.py 更改为显式 re-export
# SPDX-License-Identifier: Apache-2.0
from .vmoba import (
    moba_attn_varlen as moba_attn_varlen,
    process_moba_input as process_moba_input,
    process_moba_output as process_moba_output,
)

__all__ = ["moba_attn_varlen", "process_moba_input", "process_moba_output"]
```

评论区精华

审核人直接批准，没有提出讨论点或争议。

- 暂无高价值评论线程

风险与影响

- 风险：本 PR 仅删除未使用的代码和注释，不改变任何运行时逻辑。通过 ruff 检查（F401/F841/F821/F541/UP037）和 py_compile 编译验证，可以确保没有误删有效代码。风险极低。
- 影响：对最终用户无功能影响；对开发团队而言，减少了约 108 行待维护代码，降低了阅读和理解成本。影响范围限定在 slang/multimodal_gen 模块的 diffusion 相关子模块。
- 风险标记：低风险清理，仅删除死代码

关联脉络

- 暂无明显关联 PR