

PR #33367 完整报告

sgl-project/sglang

fix: pi05 models does not apply scale factor for language embeddings

合并时间: 2026-08-04 11:25

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/33367>

执行摘要

- 一句话: 修复 Pi0.5 语言嵌入缩放缺失, 对齐 OpenPI 输出
- 推荐动作: 值得精读。重点看 `build_openpi_model_inputs` 与 `_preprocess_image` 两个设计点: 前者演示了如何用参考实现自身的输入变换构造一致性对比, 后者用显式标记避免图像归一化歧义。该 PR 为「与官方实现的数值一致性」提供了可执行、可回归的测试范本。

功能与动机

官方 OpenPI PyTorch Pi0.5 实现 (`src/openpi/models_pytorch/pi0_pytorch.py` 的 `embed_prefix`) 在拼接图像嵌入前将语言 token 嵌入乘以 `sqrt(hidden_dim)`, SGLang 遗漏该缩放, 导致 action 输出与官方存在系统性偏差。PR body 指出原有 CI 未发现该问题: `pi05_action_http` 使用的旧 GT 由旧版 SGLang 生成, Python 一致性测试只检查可重复性与缓存复用, 手动 benchmark 又对比了经过机器人后处理的 action, 均掩盖了模型空间的差异。

实现拆解

- 在 `python/sglang/multimodal_gen/runtime/models/vlas/pi05_core.py` 的 `Pi05CoreModel.embed_prefix` 中, 对 `embed_language_tokens` 的输出补乘 `math.sqrt(lang_emb.shape[-1])`, 对齐 OpenPI 官方语义。改动位于前缀构建路径, 不触达 denoise 循环。
- `pi05_preprocess.py` 的 `_tensor_from_image` 返回值从 `torch.Tensor` 改为 `(torch.Tensor, bool, bool)`, 携带「是否字节缩放、是否已归一化」两个标记。
- `_resize_with_pad_image_tensor` 新增 `round_to_uint8` 与 `pad_value` 参数, 字节图 resize 后先 round 再映射到 `[-1, 1]`。
- 新增 `_preprocess_image` 统一入口: 已归一化的 float32 图像只做 clamp 并保留 -1 letterbox 填充, 避免二次归一化。
- `bench_pi05_openpi.py` 新增 `build_openpi_model_inputs`, 调用 OpenPI policy 的 `_input_transform` 得到与官方完全相同的模型输入 (含 token、token mask、image mask、camera order)。
- 拆出 `_openpi_model_batch` / `_openpi_model_actions`, 直接获取模型空间 action 与 SGLang 对比, 绕过 OpenPI 的机器人 unnormalization 与 horizon slicing。
- `--deterministic-noise` 模式下用 OpenPI 变换后的输入喂给 SGLang; 新增 `--action-max-abs-diff` / `--action-mean-abs-diff` 阈值参数; aloha 的 state 不再 pad 到

32 维，直接使用 14 维真实状态。

- `test_pi05_runtime_helpers.py` 新增 3 个单测：语言嵌入缩放、uint8 resize 先取整、归一化浮点图不二次归一化。
- `test_pi05_e2e.py` 将 `test_pi05_python_action_consistency` 重命名为 `test_pi05_python_action_repeatability_and_cache`，明确可重复性语义，state 默认维度回归 14。
- `testcase_configs.py` 的 CI 用例 `state_dim` 从 32 改为 14。
- `test_utils.py` 将 `SGL_TEST_FILES_CI_DATA_REVISION` 固定到 `d05810e3ea3eff1d137dec723f6e66d9c11b470f`，使一致性 CI 基于官方 provenance-backed GT。
- `docs/cookbook/vla/OpenPI/Pi0.5.mdx` 更新官方 parity 数据与对比方法说明。

关键文件：

- `python/sglang/multimodal_gen/runtime/models/vlas/pi05_core.py`（模块 模型核心；类别 source；类型 core-logic；符号 `embed_prefix`）：核心修复点：`embed_prefix` 中补上缺失的 `sqrt(hidden_dim)` 语言嵌入缩放，一行改动影响所有 Pi0.5 输出。
- `python/sglang/multimodal_gen/runtime/pipelines_core/stages/model_specific_stages/pi05_preprocess.py`（模块 预处理器；类别 source；类型 data-contract；符号 `_tensor_from_image`, `_preprocess_image`, `_resize_with_pad_image_tensor`）：图像预处理契约重构：区分字节图与已归一化浮点图，匹配 OpenPI 的 `resize` 取整与 `letterbox` 填充行为。
- `python/sglang/multimodal_gen/benchmarks/bench_pi05_openpi.py`（模块 基准脚本；类别 source；类型 dependency-wiring；符号 `build_openpi_model_inputs`, `_openpi_direct_batch`, `_openpi_model_batch`, `_openpi_model_actions`）：手动对比基准大幅升级：新增 OpenPI 模型输入构造与归一化模型空间对比，是验证修复效果的基石工具。
- `python/sglang/multimodal_gen/test/unit/test_pi05_runtime_helpers.py`（模块 单元测试；类别 test；类型 test-coverage；符号 `test_prefix_language_embedding_matches_openpi_scale`, `test_uint8_resize_rounds_before_normalization`, `test_normalized_float_image_is_not_normalized_twice`）：新增 3 个针对性单测，锁定缩放、uint8 取整与归一化不做二次映射的行为。
- `python/sglang/multimodal_gen/test/single_test_file/test_pi05_e2e.py`（模块 端到端测试；类别 test；类型 test-coverage；符号 `test_pi05_python_action_repeatability_and_cache`）：一致性测试改名落地 `repeatability` 语义，state 默认维度从 32 回归 14，与真实 Aloha 状态一致。
- `python/sglang/multimodal_gen/test/server/testcase_configs.py`（模块 测试配置；类别 test；类型 configuration）：CI 用例 `state_dim` 从 32 改为 14，使服务端一致性用例与真实 Aloha 观测一致。
- `python/sglang/multimodal_gen/test/test_utils.py`（模块 测试工具；类别 test；类型 configuration）：固定官方 GT 的 `ci-data-diffusion revision`，使一致性 CI 依赖可复现的 provenance-backed 数据。

- docs/cookbook/vla/OpenPI/Pi0.5.mdx (模块 文档; 类别 docs; 类型 documentation) :
更新官方 parity 数据与对比方法说明, 记录归一化模型空间的验证流程。

关键符号: Pi05CoreModel.embed_prefix, _preprocess_image,
_resize_with_pad_image_tensor, _tensor_from_image, build_openpi_model_inputs,
_openpi_model_batch, _openpi_model_actions

关键源码片段

python/sglang/multimodal_gen/runtime/models/vlas/pi05_core.py

核心修复点: `embed_prefix` 中补上缺失的 `sqrt(hidden_dim)` 语言嵌入缩放, 一行改动影响所有 Pi0.5 输出。

```
def embed_prefix(
    self,
    images: list[torch.Tensor],
    image_masks: list[torch.Tensor],
    tokens: torch.Tensor,
    token_masks: torch.Tensor,
) -> tuple[torch.Tensor, torch.Tensor, torch.Tensor]:
    embs = []
    pad_masks = []
    att_masks = []
    image_embs = self.paligemma_with_expert.embed_images(images)
    for image_emb, image_mask in zip(image_embs, image_masks, strict=True):
        batch_size, num_image_embs = image_emb.shape[:2]
        embs.append(image_emb)
        pad_masks.append(image_mask[:, None].expand(batch_size, num_image_embs))
        att_masks += [0] * num_image_embs

    lang_emb = self.paligemma_with_expert.embed_language_tokens(tokens)
    # 关键修复: OpenPI 官方在拼接图像嵌入前对语言嵌入乘以 sqrt(hidden_dim)。
    # 此前 SGLang 遗漏该缩放, 导致语言分支与图像分支幅度失配,
    # action 输出相对官方实现出现系统性偏差。
    lang_emb = lang_emb * math.sqrt(lang_emb.shape[-1])
    embs.append(lang_emb)
    pad_masks.append(token_masks)
    att_masks += [0] * lang_emb.shape[1]

    embs = torch.cat(embs, dim=1)
    pad_masks = torch.cat(pad_masks, dim=1)
    att_masks_t = torch.tensor(att_masks, dtype=torch.bool, device=pad_masks.device)
    att_masks_t = att_masks_t[None, :].expand(pad_masks.shape[0], len(att_masks))
    return embs, pad_masks, att_masks_t
```

python/sglang/multimodal_gen/runtime/pipelines_core/stages/model_specific_stages/pi05_preprocess.py

图像预处理契约重构：区分字节图与已归一化浮点图，匹配 OpenPI 的 resize 取整与 letterbox 填充行为。

```
def _preprocess_image(value: Any, size: tuple[int, int]) -> torch.Tensor:
    # _tensor_from_image 返回 ( 张量, 是否字节缩放, 是否已归一化到 [-1, 1])
    tensor, is_byte_scaled, is_normalized = _tensor_from_image(value)
    tensor = _resize_with_pad_image_tensor(
        tensor,
        size,
        round_to_uint8=is_byte_scaled,
        pad_value=-1.0 if is_normalized else 0.0,
    )
    if is_normalized:
        # 已归一化的 float32 图像: 只 clamp, 避免二次 2*x-1 映射
        return tensor.clamp_(-1.0, 1.0)
    return tensor * 2.0 - 1.0
```

```
def _resize_with_pad_image_tensor(
    tensor: torch.Tensor,
    size: tuple[int, int],
    *,
    round_to_uint8: bool = False,
    pad_value: float = 0.0,
) -> torch.Tensor:
    height, width = size
    if tensor.shape[-2:] == (height, width):
        return tensor
    _, cur_height, cur_width = tensor.shape
    ratio = max(cur_width / width, cur_height / height)
    resized_height = int(cur_height / ratio)
    resized_width = int(cur_width / ratio)
    tensor = F.interpolate(
        tensor[None],
        size=(resized_height, resized_width),
        mode="bilinear",
        align_corners=False,
    )[0]
    if round_to_uint8:
        # 与 OpenPI 对齐: 字节图 resize 后先 round 再映射到 [-1, 1]
        tensor = torch.round(tensor * 255.0).clamp_(0.0, 255.0) / 255.0
    pad_h0, rem_h = divmod(height - resized_height, 2)
    pad_w0, rem_w = divmod(width - resized_width, 2)
    # 归一化图用 -1 填充 letterbox, 字节图用 0 填充后再做 2*x-1
    return F.pad(
        tensor,
        (pad_w0, pad_w0 + rem_w, pad_h0, pad_h0 + rem_h),
        mode="constant",
        value=pad_value,
```

)

评论区精华

本 PR 没有实质性 review 讨论：mickqian 直接 approve，唯一的 bot 评论是 Gemini Code Assist 停服说明；另一条评论仅触发 `tag-and-rerun-ci`。关键的验收证据链在 PR body 中：官方 GT 由 OpenPI revision `15a9616...` 生成两次、SHA256 完全一致 (`319c1a1e...`)；rebased head 在 RTX 5090 直连 GPU 流式下 (13.5 GiB / 1.81 s) 通过同一一致性用例。

- 旧 CI 为何未拦截语言嵌入缩放缺失 (correctness): 引入 ci-data-diffusion#1 的官方 GT (OpenPI revision 15a9616、LeRobot checkpoint b211f3d)，并改为在归一化模型空间、相同输入与确定性噪声下对比。

风险与影响

- 风险：
 - 输出行为变更：embed_prefix 的缩放会改变所有 Pi0.5 请求的 prefix 嵌入与最终 action，依赖旧输出的用户需重新验证。
 - 预处理契约变更：_tensor_from_image 返回类型从 torch.Tensor 变为三元组，虽有单测覆盖，但其他潜在调用方需同步适配。
 - GT 依赖固定 revision：test_utils.py 固定 ci-data-diffusion 到 d05810e，若上游 OpenPI 实现或 LeRobot checkpoint 再变化，需重新生成 GT，否则 CI 会开始失败。
 - 无性能回归：缩放是 prefix 构建期的一次逐元素乘法，不进入 denoise 循环。
- 影响：
 - 用户：所有基于 Pi0.5 的 VLA 输出 action 数值会整体变化（向官方实现对齐），属于预期内的正确性修复。
 - 系统：影响面集中在 multimodal_gen 的 Pi0.5 模型族，图像预处理路径的返回契约变化需要模块内同步。
 - 团队：确立了「官方模型空间对比 + 可复现 GT provenance」的验证范式，后续 VLA 模型接入可复用 build_openpi_model_inputs 模式。
 - 风险标记：核心推理路径变更，图像预处理契约变更，CI GT 依赖固定 revision，旧行为输出不兼容

关联脉络

- PR #33317 [diffusion] Fix component accuracy topology reuse: 同为 diffusion 测试正确性修复，本 PR 在该方向基础上补充了模型空间对齐验证。
- PR #33281 [CI] Add MiniMax-H3 2-GPU consistency coverage: 建立了 diffusion 一致性 CI 机制，本 PR 在 Pi0.5 上落地官方 GT 回归。
- PR #33453 [diffusion] Restrict request-level quality to two validated tiers: lossless (default) and high: 同属 multimodal_gen 配置与文档演进，扩大 diffusion 模型支持面。