PR #26257 完整报告

sgl-project/sglang

[XPU] Fix Device Assignment

合并时间：2026-05-29 09:38

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26257

执行摘要

- 一句话：修复 XPU 设备分配，适配多模型
- 推荐动作：建议合并。该 PR 解决了 XPU 上的关键阻塞问题，设计简洁，改动量小。但建议作者或团队后续补充针对这些模型的 XPU 单元测试，并跟进 _match_cos_sin_cache_dtype 是否有更优实现（如初始化时就匹配 dtype）。

功能与动机

PR 标题和描述明确指出需要修复 XPU 上中间张量的设备分配问题，影响多个多模态模型。审查中作者进一步说明 MiniCPM-2B-128k 的 MiniCPMAttention 在 float32 下运行，而 cos_sin_cache 初始化为 bfloat16，导致 XPU 上 dtype 不匹配错误。

实现拆解

1. 导入通用设备工具函数：在 kimi_vl_moonvit.py、minicpmv.py、transformers.py、minicpmo.py 中引入 from sglang.srt.utils import get_device，替代之前的局部硬编码。
2. 替换硬编码设备字符串：将所有出现 device="cuda" 的地方（如 Rope2DPosEmb.__init__、init_resampler、init_merger、_init_parameters）改为 device=get_device()，使得张量会自动分配到当前激活的设备（XPU 或 CUDA）。
3. 修复 rotary embedding dtype 不匹配：在 rotary_embedding/base.py 的 forward_xpu 中，调用 self._match_cos_sin_cache_dtype(query) 确保 cos_sin_cache 与输入 query 的 dtype 一致，避免 XPU 上因精度不同导致的运行时错误。
4. 移除废弃的 modality 键：在 transformers_auto.py 的 _build_mm_items 中，删除 Modality.MULTI_IMAGES 条目，因为上游 Modality 枚举已不再支持该值，该删除保持与上游一致。

关键文件：

- python/sglang/srt/models/kimi_vl_moonvit.py（模块 视觉模型；类别 source；类型 data-contract；符号 Rope2DPosEmb, get_device）：核心模型文件，修改了 Rope2DPosEmb 的 __init__设备参数，并导入 get_device，直接影响 Kimi-VL 模型在 XPU 上的设备分配。
- python/sglang/srt/layers/rotary_embedding/base.py（模块 旋转嵌入；类别 source；类型 core-logic；符号 XRotaryEmbedding.forward_xpu）：新增 _match_cos_sin_cache_dtype 调用修复 XPU 上 dtype 不匹配错误，是性能敏感区域，且

审查中展开了深入讨论。

- python/sglang/srt/models/minicpmv.py（模块 视觉语言模型；类别 source；类型 data-contract；符号 MiniCPMBaseModel.init_resampler, MiniCPMBaseModel.init_merger, get_device）：包含多处 init_resampler 和 init_merger 的设备修复，影响 MiniCPM-V 系列模型在 XPU 上的运行。
- python/sglang/srt/models/transformers.py（模块 通用模型；类别 source；类型 data-contract；符号 TransformersModel._init_parameters, get_device）：修改了通用 _init_parameters 方法，影响所有通过 transformers 加载的模型在 XPU 上的参数初始化。
- python/sglang/srt/models/minicpmo.py（模块 音频模型；类别 source；类型 data-contract；符号 MiniCPMOModel.init_resampler, get_device）：Audio 模态的 init_resampler 设备修复，影响 MiniCPM 音频模型。
- python/sglang/srt/multimodal/processors/transformers_auto.py（模块 多模态处理器；类别 source；类型 core-logic；符号 TransformersAutoProcessor._build_mm_items）：移除已废弃的 Modality.MULTI_IMAGES 键，与上游 Modality 枚举变更对齐，不影响功能但需注意依赖。

关键符号：Rope2DPosEmb.init, MiniCPMBaseModel.init_resampler, MiniCPMBaseModel.init_merger, TransformersModel._init_parameters, MiniCPMOModel.init_resampler, TransformersAutoProcessor._build_mm_items, XRotaryEmbedding.forward_xpu

关键源码片段

python/sglang/srt/models/kimi_vl_moonvit.py

核心模型文件，修改了 Rope2DPosEmb 的 __init__设备参数，并导入 get_device，直接影响 Kimi-VL 模型在 XPU 上的设备分配。

```
# kimi_vl_moonvit.py (head 版本 )
from sglang.srt.utils import add_prefix, get_device # 新增导入 get_device

class Rope2DPosEmb(nn.Module):
    """2D rotary position embedding with multi-resolution support."""

    def __init__(
        self,
        dim: int,
        max_height: int,
        max_width: int,
        theta_base=10000,
        device=None, # 默认值从 'cuda' 变为 None
    ):
        super().__init__()
        self.dim = dim
        assert self.dim % 4 == 0, "dim must be divisible by 4"
        self.max_height = max_height
        self.max_width = max_width
```

```
        self.theta_base = theta_base
        # 如果调用者未指定 device，则自动获取当前平台默认设备（XPU / CUDA）
        self.device = device if device is not None else get_device()
```

python/sglang/srt/layers/rotary_embedding/base.py

新增 _match_cos_sin_cache_dtype 调用修复 XPU 上 dtype 不匹配错误，是性能敏感区域，且审查中展开了深入讨论。

```
# rotary_embedding/base.py (head 版本 )
def forward_xpu(
    self,
    query: torch.Tensor,
    key: torch.Tensor,
    positions: torch.Tensor,
    offsets: Optional[torch.Tensor] = None,
):
    """XPU 专用 forward，使用 sgl_kernel.rotary_embedding。"""
    assert self.fused_set_kv_buffer_arg is not None, (
        "fused_set_kv_buffer_arg is not supported for xpu implementation"
    )
    positions = (
        torch.add(positions, offsets) if offsets is not None else positions
    )
    # 确保 cos_sin_cache 与输入 query 的 dtype 匹配，
    # 避免 XPU 上因 float32 / bfloat16 不一致导致的运行时错误
    self._match_cos_sin_cache_dtype(query)
    return torch.ops.sgl_kernel.rotary_embedding(
        positions, query, key, self._cos_sin_cache, self.fused_set_kv_buffer_arg
    )
```

评论区精华

1. 关于 rotary embedding 的 dtype 匹配：审查者 polisettyvarma 询问是否遇到了错误，作者 SKRohit 确认 cos_sin_cache 的 dtype 与 query 不同（前者 bfloat16，后者 float32），导致错误。后来审查者 mingfeima 担心 _match_cos_sin_cache_dtype 可能带来拷贝开销，建议探讨能否从根本上避免拷贝。作者回应这是必要的，因为注意力层以 float32 运行而缓存为 bfloat16。
2. 关于 transformers_auto.py 中移除 MULTI_IMAGES：审查者 polisettyvarma 询问删除原因，作者解释 Modality.MULTI_IMAGES 已从上游 Modality 枚举中移除；后续 mingfeima 要求关联 PR 链接，作者提供了 PR #21899。

- rotary embedding 中 _match_cos_sin_cache_dtype 的必要性和性能影响 (performance)：决定保留 _match_cos_sin_cache_dtype 调用，因为它间接地解决了正确性问题，且开销在可接受范围内。
- transformers_auto.py 中移除 Modality.MULTI_IMAGES (question)：删除是合理的，与上游保持同步。

风险与影响

- 风险：
  1. 回归风险：修改涉及 6 个源文件，影响多个多模态模型。尽管 get_device() 在所有平台均可工作，但缺乏对应的单元测试，可能在其他硬件平台（如 AMD、NPU）引入意外行为。
  2. 性能风险：rotary_embedding/base.py 中的 _match_cos_sin_cache_dtype 会为每个 forward 调用执行 dtype 转换，可能引入微小开销，但通常远小于通讯开销。
  3. 废弃键移除连锁反应：移除 MULTI_IMAGES 可能影响依赖该键的下游逻辑（如某些自定义 processor），但上游维护者已同意移除。
  4. 代码质量：部分替换（如 minicpmv.py 中多处）未能统一提取共性，但改动量小，风险可控。 - 影响：影响范围：XPU 用户使用 Kimi-VL-A3B-Thinking-2506、MiniCPM-2B-128k、MiniCPM-V-2_6、llava-v1.6-vicuna-13b-hf 等模型时，以前会在设备分配阶段崩溃，现在能正常推理。对其他平台（CUDA、AMD、NPU）无功能影响，因为 get_device() 在这些平台上正确返回相应设备。性能影响：无明显退化，仅增加了一次可选的 dtype 匹配（通常一次 cast）。团队影响：简化了未来添加 XPU 设备的流程，不再需要逐个文件硬编码。 - 风险标记：缺少测试覆盖，影响多个多模态模型，dtype 匹配可能引入额外开销

关联脉络

- PR #21899 Remove Modality.MULTI_IMAGES support: 该 PR 移除了 Modality.MULTI_IMAGES，当前 PR 中的 transformers_auto.py 改动正是为了同步该变更。