PR #27341 完整报告

sgl-project/sglang

[MUSA] Fix LingBot World timestep

合并时间：2026-06-05 19:15

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27341

执行摘要

- 一句话：修复 MUSA 平台下 LingBot World 时间步数据类型
- 推荐动作：该 PR 是典型的平台兼容性修复，值得所有需要跨硬件类型运行的团队参考。尤其是 current_platform.is_float64_supported() 这种设计模式，可以作为未来处理类似数据类型兼容问题的通用范式。建议合入后，在 MUSA CI 中加入相关测试用例以防止回归。

功能与动机

根据 PR 描述，参考 https://docs.sglang.io/cookbook/diffusion/LingBot-World/LingBot-World，LingBot-World 能够在 4×MTT S5000 GPU 上成功运行，但原代码中对 timesteps 的数据类型强制使用了 double（即 float64），这在 MUSA 平台上不受支持，导致运行时错误。需要修改为根据平台能力动态选择数据类型。

实现拆解

1. 导入新模块：在文件 python/sglang/multimodal_gen/runtime/models/utils.py 顶部新增 from sglang.multimodal_gen.runtime.platforms import current_platform，以获取当前平台的能力信息。
2. 动态数据类型选择：在 pred_noise_to_pred_video 函数中，将原来直接将 scheduler.timesteps.double() 的写法，改为先通过 current_platform.is_float64_supported() 判断平台是否支持 float64，若支持则使用 float64，否则回退到 float32，从而兼容 MUSA 等仅支持 float32 的设备。
3. 影响范围：此改动仅影响 pred_noise_to_pred_video 函数中的数据类型选择逻辑，不改变其他函数或全局行为。改动简洁，但精准解决了平台兼容性问题。

关键文件：

- python/sglang/multimodal_gen/runtime/models/utils.py（模块 扩散模型；类别 source；类型 core-logic；符号 pred_noise_to_pred_video）：核心修复文件：修改了 pred_noise_to_pred_video 函数中 timesteps 的数据类型选择逻辑，新增平台浮点能力检测，确保在 MUSA 等不支持 float64 的硬件上正常运行。

关键符号：pred_noise_to_pred_video

关键源码片段

python/sglang/multimodal_gen/runtime/models/utils.py

核心修复文件：修改了 pred_noise_to_pred_video 函数中 timesteps 的数据类型选择逻辑，新增平台浮点能力检测，确保在 MUSA 等不支持 float64 的硬件上正常运行。

```
def pred_noise_to_pred_video(
    pred_noise: torch.Tensor,
    noise_input_latent: torch.Tensor,
    timestep: torch.Tensor,
    scheduler: Any,
):
    # ... 前面的形状处理逻辑不变 ...

    # 将数据转换为 float64 以进行精确计算
    pred_noise = pred_noise.double().to(device)
    noise_input_latent = noise_input_latent.double().to(device)
    sigmas = scheduler.sigmas.double().to(device)

    # 根据平台能力选择高精度数据类型：
    # MUSA（如 MTT S5000）可能不支持 float64，此时回退到 float32
    high_dtype = (
        torch.float64 if current_platform.is_float64_supported() else torch.float32
    )
    timesteps = scheduler.timesteps.to(high_dtype).to(device)

    timestep_id = torch.argmin(
        (timesteps.unsqueeze(0) - timestep.unsqueeze(1)).abs(), dim=1
    )
    sigma_t = sigmas[timestep_id].reshape(-1, 1, 1, 1)
    pred_video = noise_input_latent - sigma_t * pred_noise
    return pred_video.to(dtype)
```

评论区精华

该 PR 没有 review 评论，只有 mickqian 的批准（APPROVED）。变更简洁明了，无争议。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。改动仅影响 pred_noise_to_pred_video 一个函数内的一行数据类型转换代码，并确保在 float64 不支持的平台上回退到 float32，不会引入新的错误。但需注意：如果平台不支持 float64 且回退到 float32 时，可能会对数值精度有轻微影响，不过对于扩散模型的推理而言，这种精度损失通常可以忽略。此外，没有新增测试覆盖该分支，建议在 MUSA 平台实际验证。
- 影响：正向影响：使 LingBot World 能够在 MUSA 等仅支持 float32 的加速器上正常运行，扩大了硬件兼容性；对其他支持 float64 的平台无影响。 影响范围：仅涉及一个文件中的单行代码变更，无 API 或性能回归风险。

- 风险标记：缺少测试覆盖，精度敏感度

关联脉络

- PR #27297 [diffusion] Optimize LingBot Realtime transport and camera conditioning: 同属 LingBot World 相关功能改进，共享相同的 diffusion 模块和代码路径。