# PR #27437 完整报告

sgl-project/sglang

[diffusion] Fix LingBot-World crash on camera control with ulysses>1

合并时间：2026-06-06 22:02

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27437

## 执行摘要

- 一句话：修复 LingBot-World 相机控制崩溃
- 推荐动作：值得合并。这是一个短小、精确的 bugfix，修复了阻断性崩溃，且改动量小、风险可控。建议后续补充多 GPU 实时测试。

## 功能与动机

用户在实时服务 LingBot-World 模型时，使用相机控制与 --ulysses-degree 4（官方推荐的 4-GPU 布局）会导致每次生成块崩溃，错误为 RuntimeError: Inference tensors do not track version counter.。该问题源于 cam-conditioner 缓存键直接读取了推理模式张量的 _version 属性，而 warmup 阶段不触发此路径，因此只在 serving 时暴露。

## 实现拆解

1. 在 lingbot_world.py 中新增 _safe_tensor_version(tensor) 函数，返回 0（若 tensor.is_inference() 为真）或 tensor._version。
2. 在 _cam_conditioner_scale_shift 和 _prepare_cam_conditioner_scale_shifts 的两个缓存键构造处，将 c2ws_plucker_emb._version 替换为 _safe_tensor_version(c2ws_plucker_emb)。
3. 新增的辅助函数作为模块级函数放置于文件顶层，便于复用。

关键文件：

- python/sglang/multimodal_gen/runtime/models/dits/lingbot_world.py（模块 扩散模型；类别 source；类型 data-contract；符号 _safe_tensor_version）：单一改动文件，新增 _safe_tensor_version 辅助函数并替换两处缓存键中的 _version 调用，是修复的核心。

关键符号：_safe_tensor_version, _cam_conditioner_scale_shift, _prepare_cam_conditioner_scale_shifts

## 关键源码片段

### python/sglang/multimodal_gen/runtime/models/dits/lingbot_world.py

单一改动文件，新增 _safe_tensor_version 辅助函数并替换两处缓存键中的 _version 调用，是修复的核心。

```
# 在 inference_mode 下，张量不跟踪版本计数器，直接访问 _version 会抛出 RuntimeError。
```

```
# _safe_tensor_version 通过 is_inference() 判断，返回 0 作为安全的回退值。
def _safe_tensor_version(tensor: torch.Tensor) -> int:
    """Return ``tensor._version``, or ``0`` for inference-mode tensors.

    Tensors created under ``torch.inference_mode`` do not track a version
    counter, so reading ``tensor._version`` raises ``RuntimeError``. The value
    is only used as a cache-invalidation hint for the camera conditioner, so a
    constant fallback is safe for such tensors.
    """
    return 0 if tensor.is_inference() else tensor._version


class CausalLingBotWorldTransformerBlock(...):
    def _cam_conditioner_scale_shift(self, c2ws_plucker_emb):
        # ...
        source_key = (
            c2ws_plucker_emb.data_ptr(),
            tuple(c2ws_plucker_emb.shape),
            tuple(c2ws_plucker_emb.stride()),
            c2ws_plucker_emb.dtype,
            c2ws_plucker_emb.device.type,
            c2ws_plucker_emb.device.index,
            _safe_tensor_version(c2ws_plucker_emb), # 之前是 c2ws_plucker_emb._version
        )
        # 其余逻辑不变
```

# 评论区精华

Review 中无讨论，仅由 gemini-code-assist[bot] 自动审查后确认改动正确，mickqian 直接批准合并。Issue 评论中 mickqian 建议增加多 GPU 实时测试，但未纳入本次 PR。

- 暂无高价值评论线程

# 风险与影响

- 风险：风险极低。更改仅限于两处缓存键构造，对推理模式张量返回 0 是安全的（版本号仅用作缓存失效提示）。但若未来引入需要精确缓存失效的场景，常数值 0 可能导致缓存未正确失效。建议添加注释说明此假设。
- 影响：影响范围限于实时服务 LingBot-World 模型的用户，特别是使用相机控制且 ulysses_degree > 1 的场景。修复后所有 GPU 布局（ulysses=1/2/4）均可正常运行。
- 风险标记：缺少测试覆盖

# 关联脉络

- PR #27383 [diffusion] Optimize LingBot realtime SP cache path: 同样涉及 LingBot-World 实时缓存路径的优化，本 PR 修复了该缓存路径中的一个崩溃。