PR #27077 完整报告

sgl-project/sglang

[diffusion] Preserve dtype in WanVAE nearest upsample

合并时间：2026-06-03 08:32

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27077

执行摘要

- 一句话：WanVAE 上采样保持输入 dtype
- 推荐动作：小优化，可直接合并。关注点在于 current_platform.is_amp_supported() 的语义是否覆盖所有 AMP 场景。

功能与动机

WanVAE nearest 上采样始终将输入转为 fp32 再 type_as 回原 dtype，在支持 AMP 的平台（如 H200）上会引入不必要的精度转换开销。PR body 提到通过保持 dtype 在 AMP 平台可减少峰值内存（44174MB→43980MB），并保持解码速度不变。

实现拆解

1. 修改入口：在 python/sglang/multimodal_gen/runtime/models/vaes/parallel/wan_common_utils.py 的 WanUpsample.forward 方法中增加平台检测分支。
2. 核心逻辑：使用 current_platform.is_amp_supported() 判断当前平台是否支持 AMP。若支持，直接调用 super().forward(x) 保持输入 dtype；否则保留旧逻辑 super().forward(x.float()).type_as(x) 作为 fallback。
3. 没有测试、配置或部署配套变更：仅 2 行源码改动，无配套测试或配置更新。

关键文件：

- python/sglang/multimodal_gen/runtime/models/vaes/parallel/wan_common_utils.py（模块 扩散模型；类别 source；类型 core-logic；符号 WanUpsample.forward）：唯一修改文件，WanUpsample 类新增平台检测分支以保持 dtype。

关键符号：WanUpsample.forward

关键源码片段

python/sglang/multimodal_gen/runtime/models/vaes/parallel/wan_common_utils.py

唯一修改文件，WanUpsample 类新增平台检测分支以保持 dtype。

```
# python/sglang/multimodal_gen/runtime/models/vaes/parallel/wan_common_utils.py
class WanUpsample(nn.Upsample):
    r"""
```

```
    Perform upsampling while ensuring the output tensor has the same data type as the input.
    """

    def forward(self, x):
        # AMP 支持的平台（如 H200）直接上采样，保持输入 dtype，
        # 避免不必要的 fp32 转换和 type_as 开销。
        if current_platform.is_amp_supported():
            return super().forward(x)
        # 非 AMP 平台（如某些 CPU 或旧 GPU）回退到旧逻辑：
        # 转为 fp32 计算后再转换回输入 dtype。
        return super().forward(x.float()).type_as(x)
```

评论区精华

该 PR 没有 review 评论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅增加一个条件分支，不影响非 AMP 平台行为；AMP 平台上消除了 type_as 操作，可能轻微影响数值精度一致性（但 VAE 对精度不敏感）；无测试覆盖，但基准验证了输出一致性。
- 影响：影响范围极小：仅影响 WanVAE 模型在 AMP 平台的 forward 路径；性能影响中性（解码时间持平，内存微降）；无功能变更。
- 风险标记：缺少测试覆盖

关联脉络

- 暂无明显关联 PR