PR #27037 完整报告

sgl-project/sglang

[diffusion] Enable Cosmos3 parallel decode

合并时间：2026-06-02 18:18

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27037

执行摘要

- 一句话：开启 Cosmos3 VAE 并行解码
- 推荐动作：建议合并。该 PR 是低风险、高收益的小配置变更，验证充分。可精读 cosmos3.py 的 __post_init__ 部分以理解并行 VAE 在 Cosmos3 中的设计权衡。

功能与动机

在 Cosmos3 模型中，WanVAE 默认会为并行编码 / 解码启用 SP 分片路径，但并行编码会改变 I2V 的条件潜在特征导致像素错乱（PR body 描述）。因此需要在配置中禁用并行编码并启用并行解码，以在保持 I2V 正确性的同时获得多 GPU 解码加速。

实现拆解

1. 修改配置项：在 python/sglang/multimodal_gen/configs/pipeline_configs/cosmos3.py 的 Cosmos3VideoConfig.__post_init__ 方法中，将 self.vae_config.use_parallel_decode 从 False 改为 True。
2. 更新注释：同步修改相关注释，解释为何保留并行编码为 False（防止改变 I2V 条件潜在特征），并将原先描述“默认并行导致 SP 分片产生乱码像素”的注释替换为更准确的说明。
3. 未涉及测试：本次改动未新增或修改测试文件，PR body 说明使用远程 GPU 开发机验证并通过 CI。

关键文件：

- python/sglang/multimodal_gen/configs/pipeline_configs/cosmos3.py（模块 扩散配置；类别 source；类型 core-logic）：核心配置文件，通过一行配置改动启用多 GPU 并行 VAE 解码，并更新注释说明设计决策。

关键符号：未识别

关键源码片段

python/sglang/multimodal_gen/configs/pipeline_configs/cosmos3.py

核心配置文件，通过一行配置改动启用多 GPU 并行 VAE 解码，并更新注释说明设计决策。

```
# python/sglang/multimodal_gen/configs/pipeline_configs/cosmos3.py
# 在 Cosmos3VideoConfig 的 __post_init__ 方法中，
# 开启 VAE 并行解码（多 GPU 加速），并保持编码串行以避免 I2V 条件潜在特征被破坏。
```

```
def __post_init__(self):
    self.vae_config.arch_config.z_dim = 48
    # Encoder is needed for I2V; T2V/T2I never invoke it.
    self.vae_config.load_encoder = True
    self.vae_config.load_decoder = True
    # keep WanVAE encode replicated because parallel encode changes I2V
    # conditioning latents when sp_world_size > 1
    self.vae_config.use_parallel_encode = False
    self.vae_config.use_parallel_decode = True # 启用并行解码，多 GPU 分片
```

评论区精华

仅有一条来自 gemini-code-assist[bot] 的自动化评论，总结变更内容并说明未收到人工 review 反馈。无实质讨论交锋。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 回归风险低：单 GPU 运行时 sp_world_size == 1，并行解码退化为非分片路径，行为不变。多 GPU 场景下解码加速已验证，输出质量与原始接近（PSNR 40.06dB）。
  2. 潜在兼容性：若未来 WanVAE 并行解码逻辑对 Shard 数量或批次形状有隐含假设，可能在非标准配置下出错，但当前验证覆盖 4xH200。
  3. 缺少测试覆盖：无新增自动化测试，回归依赖外部验证。 - 影响：影响范围：仅 Cosmos3 模型配置，不涉及其他模型或模块。用户影响：多 GPU 用户将获得更快的 VAE 解码和更低峰值内存；单 GPU 用户无感知。系统影响：无。 - 风险标记：缺少测试覆盖

关联脉络

- PR #26973 [diffusion] reduce Cosmos3 denoise overhead: 同一模型（Cosmos3）持续优化，前序 PR 减少去噪开销，本 PR 进一步加速 VAE 解码。