PR #24022 完整报告

sgl-project/sglang

[diffusion] fix: improve LTX2.3 reference accuracy controls

合并时间：2026-04-29 21:39

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24022

执行摘要

- 一句话：改进 LTX2.3 参考精度与对齐控制
- 推荐动作：该 PR 展示了如何通过逐步对齐官方实现来提升扩散模型管线精度，特别是文本连接器 CFG 分支的重构和 V2A 跳过标记的设计，具有参考价值。对于关注扩散模型质量工程化的团队建议精读。

功能与动机

根据 PR 描述，目标是改进 LTX-2.3 one-stage TI2V 的参考对齐控制，以更紧密匹配官方行为，同时更新 CI 一致性阈值以适配重新生成的支持 V2A 的 ground truth。具体需要修正文本连接器中 CFG 分支的处理、音频跨注意力的 RoPE 缩放因子，并增加向后兼容旧版 GT 的标记。

实现拆解

1. 重构文本连接器的 CFG 分支（text_connector.py）：将原先的正负条件拼接一起送入 connector 的方式改为分别调用 connector 两次，消除批处理导致的数值差异；同时添加严格的输入非空校验。
2. 添加 V2A 交叉注意力跳过标记（ltx_2_denoising.py 和 ltx_2.py 采样参数）：在 LTX23SamplingParams 中新增 skip_v2a_cross_attn_for_video_gt 字段（默认 False），该标记通过请求 extra 传递到去噪阶段，在模型 forward 调用时注入 disable_v2a_cross_attn 参数，允许复现旧版不带 V2A 的 GT。
3. 修复音频跨注意力 RoPE（ltx_2.py DiT 模型）：在模型初始化中向 cross-attn pos embed 传入 scale_factors=self.audio_scale_factors，修正音频位置的缩放因子。
4. 保留 --disable-autocast 标志（configs/utils.py 和 test_server_args.py）：确保当通过配置字典覆盖 pipeline 配置时 disable_autocast 字段不被丢弃；配套新增单元测试验证。
5. 更新 CI 一致性阈值与性能基线（consistency_threshold.json 和 perf_baselines.json）：为多个 LTX-2.3 测试用例更新并新增阈值条目，确保 CI 通过。

关键文件：

- python/sglang/multimodal_gen/runtime/pipelines_core/stages/text_connector.py（模块 文本连接器；类别 source；类型 core-logic）：核心重构 CFG 分支，独立处理正负条件以匹配官方行为，并添加输入校验
- python/sglang/multimodal_gen/runtime/pipelines_core/stages/ltx_2_denoising.py（模块 去噪阶段；类别 source；类型 core-logic）：添加 skip_v2a_cross_attn 逻辑，控制模型

forward 中的视频 - 音频交叉注意力

- python/sglang/multimodal_gen/configs/sample/ltx_2.py（模块 采样参数；类别 source；类型 core-logic）：采样参数新增 skip_v2a_cross_attn_for_video_gt 字段，请求 extra 传递
- python/sglang/multimodal_gen/test/unit/test_server_args.py（模块 服务参数；类别 test；类型 test-coverage；符号 test_disable_autocast_is_preserved_after_pipeline_config_resolution）：新增测试确保 --disable-autocast 在 pipeline 配置解析后保留
- python/sglang/multimodal_gen/test/server/consistency_threshold.json（模块 测试阈值；类别 test；类型 test-coverage）：更新 LTX-2.3 各类测试的一致性阈值，添加新测试条目
- python/sglang/multimodal_gen/runtime/models/dits/ltx_2.py（模块 模型定义；类别 source；类型 data-contract）：修复音频 cross-attention RoPE 缩放因子传递
- python/sglang/multimodal_gen/configs/utils.py（模块 配置工具；类别 source；类型 core-logic）：修复 pipeline 配置覆盖时 disable_autocast 字段被丢弃
- python/sglang/multimodal_gen/test/server/perf_baselines.json（模块 性能基线；类别 test；类型 test-coverage）：更新性能基线值

关键符号：LTX2TextConnectorStage.forward, LTX23SamplingParams.build_request_extra, evaluate_stage1_guided_x0, test_disable_autocast_is_preserved_after_pipeline_config_resolution

关键源码片段

python/sglang/multimodal_gen/runtime/pipelines_core/stages/text_connector.py

核心重构 CFG 分支，独立处理正负条件以匹配官方行为，并添加输入校验

```
def forward(self, batch: Req, server_args: ServerArgs) -> Req:
    # Input: batch . prompt_embeds (from Gemma, [B, S, D])
    # Output: batch . prompt_embeds (Video Context), batch . audio_prompt_embeds (Audio Context)

    prompt_embeds = batch.prompt_embeds
    prompt_attention_mask = batch.prompt_attention_mask
    neg_prompt_embeds = batch.negative_prompt_embeds
    neg_prompt_attention_mask = batch.negative_attention_mask

    # 处理列表形式（从列表取第一个元素）
    if isinstance(prompt_embeds, list):
        prompt_embeds = prompt_embeds[0] if len(prompt_embeds) > 0 else None
    if isinstance(prompt_attention_mask, list):
        prompt_attention_mask = (
            prompt_attention_mask[0] if len(prompt_attention_mask) > 0 else None
        )
    if isinstance(neg_prompt_embeds, list):
        neg_prompt_embeds = (
```

```
            neg_prompt_embeds[0] if len(neg_prompt_embeds) > 0 else None
        )
    if isinstance(neg_prompt_attention_mask, list):
        neg_prompt_attention_mask = (
            neg_prompt_attention_mask[0] if len(neg_prompt_attention_mask) > 0 else None
        )

    # 添加严格输入检查：prompt_embeds 和注意力掩码不能为空
    if prompt_embeds is None or prompt_attention_mask is None:
        raise ValueError(
            "LTX2TextConnectorStage requires prompt embeddings and "
            "attention mask."
        )

    if batch.do_classifier_free_guidance:
        # CFG 模式下，negative 条件也必须存在
        if neg_prompt_embeds is None or neg_prompt_attention_mask is None:
            raise ValueError(
                "LTX2TextConnectorStage requires negative prompt embeddings "
                "and attention mask when classifier-free guidance is enabled."
            )

        # 官方 LTX-2.3 将正负条件独立送入 connector，批处理会改变输出数值
        dtype = prompt_embeds.dtype
        pos_additive_mask = (
            (prompt_attention_mask.to(torch.int64) - 1).to(dtype)
            * torch.finfo(dtype).max
        )
        neg_additive_mask = (
            (neg_prompt_attention_mask.to(torch.int64) - 1).to(dtype)
            * torch.finfo(dtype).max
        )

        with set_forward_context(current_timestep=None, attn_metadata=None):
            # 分别调用 connector 处理正负条件
            pos_embeds, pos_audio_embeds, pos_mask = self.connectors(
                prompt_embeds, pos_additive_mask, additive_mask=True
            )
            neg_embeds, neg_audio_embeds, neg_mask = self.connectors(
                neg_prompt_embeds, neg_additive_mask, additive_mask=True
            )

        # 更新 batch 中的正负嵌入和掩码
        batch.prompt_embeds = [pos_embeds]
        batch.audio_prompt_embeds = [pos_audio_embeds]
        batch.prompt_attention_mask = pos_mask
        batch.negative_prompt_embeds = [neg_embeds]
        batch.negative_audio_prompt_embeds = [neg_audio_embeds]
        batch.negative_attention_mask = neg_mask
```

```
    else:
        # 非 CFG 模式：按 Diffusers 原始实现处理
        dtype = prompt_embeds.dtype
        additive_attention_mask = (
            (prompt_attention_mask.to(torch.int64) - 1).to(dtype)
            * torch.finfo(dtype).max
        )
        with set_forward_context(current_timestep=None, attn_metadata=None):
            connector_prompt_embeds, connector_audio_prompt_embeds, connector_mask = (
                self.connectors(prompt_embeds, additive_attention_mask, additive_mask=True)
            )

        # 仅更新正字段
        batch.prompt_embeds = [connector_prompt_embeds]
        batch.audio_prompt_embeds = [connector_audio_prompt_embeds]
        batch.prompt_attention_mask = connector_mask

    return batch
```

python/sglang/multimodal_gen/runtime/pipelines_core/stages/ltx_2_denoising.py

添加 skip_v2a_cross_attn 逻辑，控制模型 forward 中的视频 - 音频交叉注意力

```
# 在 stage1 引导函数开始处从 batch.extra 中读取标记
skip_v2a_cross_attn_for_video_gt = bool(
    batch.extra.get("ltx2_skip_v2a_cross_attn_for_video_gt", False)
)

# ... 在 evaluate_stage1_guided_x0 内部调用模型时传递 disable_v2a_cross_attn
v_pos, a_v_pos = step.current_model(
    **self._build_ltx2_model_kwargs(
        ctx,
        base_model_kwargs_local,
        encoder_hidden_states=encoder_hidden_states,
        audio_encoder_hidden_states=audio_encoder_hidden_states,
        encoder_attention_mask=encoder_attention_mask,
        disable_v2a_cross_attn=(
            skip_v2a_cross_attn_for_video_gt
        ),
    )
)
# 类似的传递也出现在 v_neg、perturbed 和 modality 调用中
```

python/sglang/multimodal_gen/configs/sample/ltx_2.py

采样参数新增 skip_v2a_cross_attn_for_video_gt 字段，请求 extra 传递

```
@dataclasses.dataclass
class LTX23SamplingParams(LTX2SamplingParams):
    """Sampling parameters matching official LTX-2.3 one-stage defaults."""
```

```
    # ... 其他字段 ...
    skip_v2a_cross_attn_for_video_gt: bool = False # 新增标记，默认不跳过

    def build_request_extra(self) -> dict[str, Any]:
        extra = super().build_request_extra()
        extra["ltx2_stage1_guider_params"] = {
            # ... 原有参数 ...
        }
        # 当标记为 True 时，才将其写入 extra，否则保持兼容
        if self.skip_v2a_cross_attn_for_video_gt:
            extra["ltx2_skip_v2a_cross_attn_for_video_gt"] = True
        return extra
```

评论区精华

审查者 gemini-code-assist[bot] 提出了三点核心关注：

1. FP8 cast 的 CPU 兼容性：在 transformer_load_utils.py 中尝试将模型权重转换为 torch.float8_e4m3fn 时，若张量位于 CPU 会导致运行时崩溃。作者在后续提交中已将该项目分离至独立 PR #24024，本 PR 不再包含此代码。
2. TextConnector CFG 分支空指针：若 prompt_embeds 或 attention_mask 为 None，访问 .dtype 或调用 .to() 会引发 AttributeError。作者添加了前置 ValueError 检查，问题已解决。
3. TextConnector 非 CFG 分支同样风险：类似问题已由同一前置校验覆盖，已解决。

- FP8 cast 的 CPU 兼容性风险 (correctness): 作者将该功能分离到独立 PR #24024，本 PR 移除该代码，风险消除。
- TextConnector CFG 分支空指针检查 (correctness): 作者在 head 代码中添加了 if prompt_embeds is None 的检查，引发 ValueError。
- TextConnector 非 CFG 分支空指针检查 (correctness): 已通过前置的 prompt_embeds is None 检查覆盖。

风险与影响

- 风险：
  - 核心逻辑变更：独立调用 connector 两次会增加一次前向传播，计算量略有增加但对整体 Pipeline 影响微小。
  - 数据契约变更：新增 skip_v2a_cross_attn_for_video_gt 字段，默认值 False 确保向后兼容；仅当显式设置后才出现在 extra 中。
  - CI 阈值调整：一致性阈值基于重新生成的 GT 设定，min_ssim=0.79，阈值设置合理，但可能掩盖轻微回归；建议持续监控。
  - disable-autocast 保留修复：确保该标志在配置覆盖后仍生效，避免精度模式意外切换。
  - 无安全性或性能风险。
  - 影响：影响范围：仅限于 LTX-2.3 扩散管线（one-stage TI2V 和 two-stage T2V/TI2V），对其他模型无影响。用户影响：获得更准确的参考对齐；需要复现旧版行为的用户可设

置 skip_v2a_cross_attn_for_video_gt=True。系统影响：无。测试影响：CI 一致性测试覆盖 LTX-2.3 的阈值已更新，新增两条测试条目。

- 风险标记：数据契约变更，CI 阈值调整可能掩盖回归，管线分支增加

关联脉络

- PR #23714 [diffusion] CI: update ground truth with official output: 本 PR 的一致性阈值更新依赖于该 PR 引入的重新生成 GT。