PR #24988 完整报告

sgl-project/sglang

Fix DenoisingStage to respect dit_precision config instead of hardcoded bfloat16

合并时间：2026-05-15 13:22

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24988

执行摘要

- 一句话：修复 denoising 阶段忽略 dit_precision 配置的 bug
- 推荐动作：值得合并，修复逻辑清晰，风险极低。建议作为常规维护 PR 处理。

功能与动机

在 DenoisingStage 中，target_dtype 被硬编码为 torch.bfloat16，忽略了用户配置的 pipeline_config.dit_precision。这与其他 stage（如 helios_denoising、mova）不一致——它们正确使用 PRECISION_TO_TYPE[server_args.pipeline_config.dit_precision] 来确定目标 dtype。该 bug 意味着即使用户显式设置 dit_precision 为 fp16 或 fp32，denoising stage 仍会强制将 tensor 转换为 bfloat16，导致精度损失，尤其对以 fp16 或 fp32 训练或微调的模型可能降低生成质量。

实现拆解

变更仅涉及一行代码：

1. 在文件 python/sglang/multimodal_gen/runtime/pipelines_core/stages/denoising.py 第 621 行，将 target_dtype = torch.bfloat16 替换为 target_dtype = PRECISION_TO_TYPE[server_args.pipeline_config.dit_precision]。

该行位于 _prepare_denoising_loop 方法中，负责设置精度和自动类型转换。修改后，denoising stage 会使用用户配置的 dit_precision，与 helios_denoising、mova 等其他 stage 保持一致。autocast_enabled 逻辑维持不变。

由于仅替换常量，无需额外测试或配置变更，不影响现有逻辑。

关键文件：

- python/sglang/multimodal_gen/runtime/pipelines_core/stages/denoising.py（模块 扩散模型；类别 source；类型 core-logic；符号 _prepare_denoising_loop）：核心变更文件，修复了 DenoisingStage 中硬编码 bfloat16 的问题，改为动态读取 dit_precision 配置。

关键符号：_prepare_denoising_loop

关键源码片段

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

核心变更文件，修复了 DenoisingStage 中硬编码 bfloat16 的问题，改为动态读取 dit_precision 配置。

```
# 文件：python/sglang/multimodal_gen/runtime/pipelines_core/stages/denoising.py
# 方法：_prepare_denoising_loop

# 设置精度和自动类型转换
# 修复前：target_dtype = torch.bfloat16（忽略 dit_precision 配置）
# 修复后：使用用户配置的 dit_precision，与其他 stage 保持一致
target_dtype = PRECISION_TO_TYPE[server_args.pipeline_config.dit_precision]

autocast_enabled = (
    target_dtype != torch.float32
) and not server_args.disable_autocast
```

评论区精华

无 review 评论。合并者 mickqian 直接批准了 PR。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅将硬编码常量改为配置读取，且该配置在其他 stage 中已被广泛使用。当 dit_precision 为 bfloat16（常见默认）时行为完全不变。
- 影响：影响范围：仅影响 DenoisingStage 的精度选择。用户若设置 dit_precision 为 fp16 或 fp32，将正确使用所需精度，避免意外精度截断。不影响性能。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR