

PR #29041 完整报告

sgl-project/sglang

[diffusion] fix: paint multiview VAE must follow UNet dtype

合并时间: 2026-06-24 23:54

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29041>

执行摘要

- 一句话: 修复 Hunyuan3D 纹理 VAE dtype 与 UNet 不匹配
- 推荐动作: 值得快速合入, 修复明确且安全。建议开发者关注类似 VAE/DiT dtype 独立配置模式是否存在其他管线中存在同样风险。

功能与动机

HunyuanPaint 纹理生成管线因 VAE 与 UNet 的 dtype 不一致导致崩溃, 且该崩溃被异常处理遮蔽, 使得运行看似成功但输出纹理错误。CI 中 `hunyuan3d_shape_gen` 测试禁用了 `paint` 路径, 因此问题未被发现。PR body 明确说明官方管线 VAE 与 UNet 均在 fp16 下运行, `vae_precision` 配置本用于 3D ShapeVAE, 不应影响纹理 VAE。

实现拆解

1. 移除纹理 VAE 的独立 dtype 解析: 在 `paint.py` 的 `_do_load_paint` 方法中, 删除原先通过 `PRECISION_TO_TYPE.get(getattr(self.config, "vae_precision", "fp32"), torch.float32)` 获取 `vae_dtype` 的代码段。
2. 保持 UNet dtype 解析逻辑不变: 继续从 `dit_precision` 配置解析 `dit_dtype`, 并保留 CPU/MPS 设备上的 float32 降级逻辑。
3. 将纹理 VAE 的 dtype 直接设为 UNet dtype: 新增 `vae_dtype = dit_dtype` 一行赋值, 确保纹理 VAE 的精度与 UNet 完全一致。
4. 更新注释: 添加多行注释解释该变更的原因, 说明参考注意力的数据流以及 `vae_precision` 的作用范围。

关键文件:

- `python/sglang/multimodal_gen/runtime/pipelines_core/stages/model_specific_stages/hunyuan3d/paint.py` (模块 扩散管线; 类别 source; 类型 data-contract; 符号 `_do_load_paint`): 核心变更文件, 修改了纹理 VAE 的 dtype 解析逻辑, 将 `vae_dtype` 设为 `dit_dtype` 而非独立读取 `vae_precision` 配置。

关键符号: `_do_load_paint`

关键源码片段

python/sglang/multimodal_gen/runtime/pipelines_core/stages/model_specific_stages/hunyuan3d/paint.py

核心变更文件，修改了纹理 VAE 的 dtype 解析逻辑，将 vae_dtype 设为 dit_dtype 而非独立读取 vae_precision 配置。

```
def _do_load_paint(self, server_args: ServerArgs) -> None:
    # ... ( 前面的 VAE 权重加载代码不变 ) ...
    self.vae.load_state_dict(state_dict)

    # Resolve the DiT (multiview UNet) dtype from config, with CPU/MPS fallback.
    dit_dtype = PRECISION_TO_TYPE.get(
        getattr(self.config, "dit_precision", "fp16"), torch.float16
    )
    if self.device.type in ("cpu", "mps") and dit_dtype in (
        torch.float16,
        torch.bfloat16,
    ):
        dit_dtype = torch.float32
    # The multiview (Stable-Diffusion) AutoencoderKL must share the UNet dtype.
    # Reference attention feeds its VAE-encoded ref_latents straight into the
    # fp16 UNet, and the official HunyuanPaint pipeline runs VAE+UNet entirely
    # in fp16. The `vae_precision` knob targets the 3D ShapeVAE (geometry
    # precision) — applying it to this 2D texture VAE produces an
    # fp32-input / fp16-weight mismatch that crashes the paint UNet.
    vae_dtype = dit_dtype # <-- 核心变更：不再从 vae_precision 配置读取

    self.vae = self.vae.to(device=self.device, dtype=vae_dtype).eval()
    # ... ( 后续 UNet 和 scheduler 加载不变 ) ...
```

评论区精华

无实质性 review 讨论。仅有 Gemini Code Assist 的自动代码审查评论，确认了变更内容且无额外反馈。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅影响 Hunyuan3D Paint 纹理生成路径中的 VAE 加载部分，且仅在模型加载时执行一次。纹理 VAE 从 fp32 改为 fp16 可能会带来微小的数值精度损失，但官方管线即运行于 fp16，且该 VAE 编码的结果用于纹理生成而非对精度敏感的场景。其他 10 个 VAE 编码点已确认安全，无副作用。
- 影响：直接影响 Hunyuan3D Paint 纹理生成管线，修复了因 dtype 不匹配导致的完全失败和视觉错误。间接影响微乎其微，因为只涉及一个文件的一行核心赋值变更。对 3D 形状生成路径无影响。
- 风险标记：缺少测试覆盖

关联脉络

- PR #28832 [diffusion] Fix Qwen-Image-Layered latent shape: 同为 diffusion 模块的 dtype/ 形状相关 bugfix, 涉及 VAE 与 UNet 的精度配合。
- PR #27088 Add align_tensor_to_module_dtype utility: PR body 中提及此工具, 说明本 PR 选择不依赖对齐辅助函数, 而是从源头统一 dtype。
- PR #28781 [Ci related?] Add paint_enable=false config for hunyuan3d_shape_gen test: PR body 指出该 PR 在 CI 中添加了 paint_enable=false, 导致 paint 纹理生成路径未被测试覆盖, 间接使本 bug 未被发现。