PR #24358 完整报告

sgl-project/sglang

[Codex] Diffusion tune Hunyuan3D shape export chunks

合并时间：2026-05-15 15:31

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24358

执行摘要

- 一句话：调整 Hunyuan3D shape 导出块大小以提升性能
- 推荐动作：PR 变更简单清晰，基于 profile 数据驱动，收益明确且风险可控，值得合并。对于显存敏感的场景，可考虑保留可配置性。

功能与动机

在 Hunyuan3D denoise 融合工作完成后，shape-only 路径的耗时集中在 Hunyuan3DShapeExportStage 和 Hunyuan3DShapeSaveStage。导出阶段仍有许多小的 VAE/geo-decoder 块启动，调整块大小可减少启动开销。

实现拆解

1. 在 python/sglang/multimodal_gen/configs/pipeline_configs/hunyuan3d.py 的 Hunyuan3D2PipelineConfig 数据类中，将 shape_num_chunks 字段的默认值从 8000 改为 32000。
2. 这是该 PR 唯一的代码变更，未涉及其他逻辑、测试或配置文件修改。
3. 通过验证确保输出 mesh 的哈希值与基线一致，保证正确性。

关键文件：

- python/sglang/multimodal_gen/configs/pipeline_configs/hunyuan3d.py（模块 扩散模型；类别 source；类型 core-logic；符号 Hunyuan3D2PipelineConfig.shape_num_chunks）：核心变更文件，修改了 shape_num_chunks 默认值，直接影响 shape 导出阶段的块大小和性能。

关键符号：未识别

关键源码片段

python/sglang/multimodal_gen/configs/pipeline_configs/hunyuan3d.py

核心变更文件，修改了 shape_num_chunks 默认值，直接影响 shape 导出阶段的块大小和性能。

```
@dataclass
class Hunyuan3D2PipelineConfig(PipelineConfig):
    # ... 其他字段不变 ...
```

```
    # Shape 模型配置
    shape_model_path: Optional[str] = None
    shape_use_safetensors: bool = True
    shape_variant: Optional[str] = "fp16"
    shape_num_inference_steps: int = 50
    guidance_scale: float = 5.0
    shape_box_v: float = 1.01
    shape_octree_resolution: int = 384
    shape_mc_level: float = 0.0
    shape_mc_algo: Optional[str] = "mc"
    # 块大小从 8000 调整为 32000，减少 VAE/geo-decoder 启动次数
    shape_num_chunks: int = 32000
    shape_output_type: str = "trimesh"
    # ... 后续字段不变 ...
```

评论区精华

只有 gemini-code-assist[bot] 的自动评论指出变更内容，未产生实质性讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。峰值显存增加 408 MB（约 6.4%），在 H100 上从 6430 MB 升至 6838 MB，仍远低于显存上限。如果其他场景显存压力较大，此默认值可能不适用。未增加测试覆盖，建议后续添加块大小相关的回归测试。
- 影响：影响范围限于 Hunyuan3D 模型的 shape 生成流程。默认块大小变更对所有使用该配置的用户生效，但用户可通过显式设置 shape_num_chunks 覆盖。性能提升约 5-8%，显存增加约 6%。
- 风险标记：缺少测试覆盖，显存增加约 6%

关联脉络

- PR #25305 [diffusion] Fix Z-Image Cache-DiT sequence-parallel override: 同为 diffusion 模块的 bugfix，涉及模型配置调整。
- PR #24988 Fix DenoisingStage to respect dit_precision config instead of hardcoded bfloat16: 同为 diffusion 模块的配置修正，与本 PR 的配置调优属于同一功能方向。