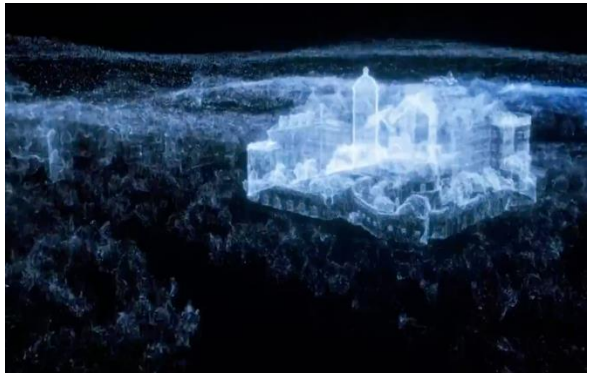
高效可控的三维场景生成

章国锋

浙江大学计算机辅助设计与图形系统全国重点实验室

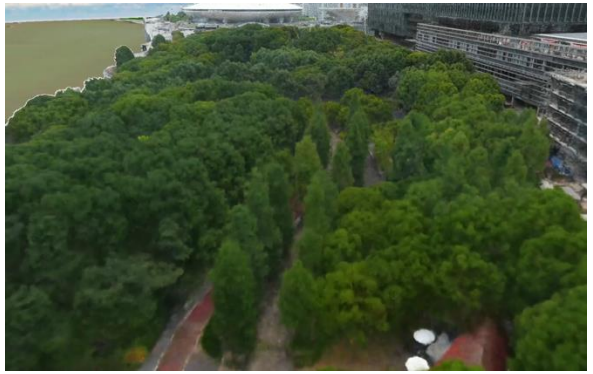
研究背景与意义

对真实场景三维数字化以及生成逼真的虚拟三维场景，而且可编辑和交互，具有重要的研究价值，可以广泛应用于AR/VR、元宇宙、智慧文旅、自动驾驶、游戏、仿真数据生成等



华为Petal Maps 3D实景地图渲染

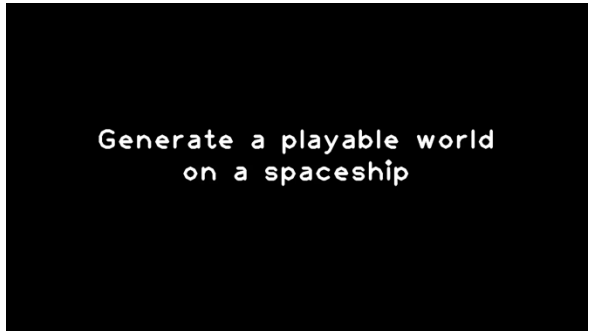
Google Map基于云串流神经渲染的三维城市预览



商汤琼宇平台

黑神话 · 悟空

World Labs 场景生成



DeepMind Genie2 游戏生成

Street Gaussians 自动驾驶场景重建

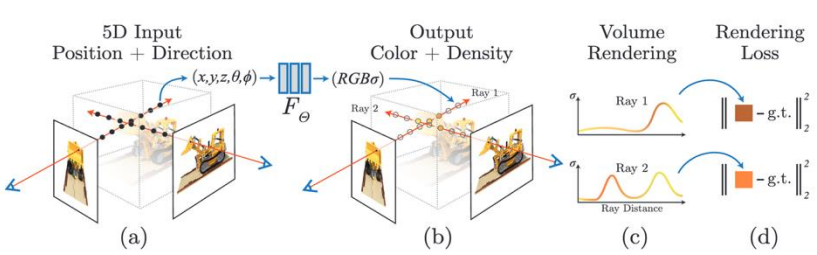
研究背景与意义

传统视觉三维重建

- **显式几何纹理**表达
  - 精细化建模有难度
  - 难以处理高反光、半透明

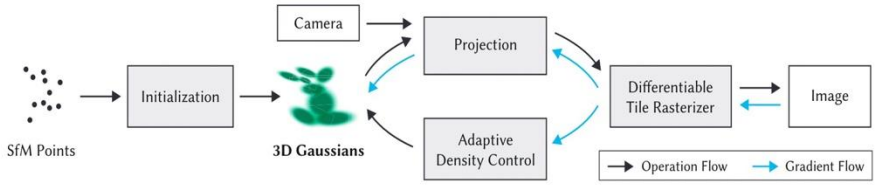
基于隐式神经表示的三维重建

- **隐式神经**表达
  - 利于表达精细结构，支持高反光、半透明
  - 重建、渲染耗时耗算力



基于三维高斯表达的三维重建

- **显式3DGS**表达
  - 支持精细化建模、轻量化实时渲染
  - 大规模场景可拓展性高
  - 需要采集稠密图像



可微3D高斯表达与Splatting方法

3DGS大规模场景重建与
Web 2K实时渲染

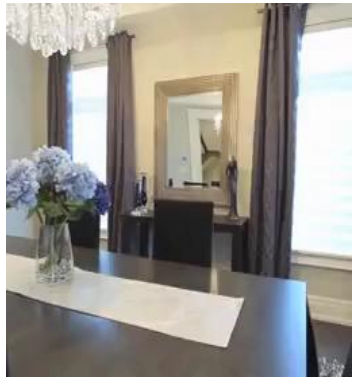
- 探索利用**更稀疏**的图像实现**高效、完整、高质量**的三维场景重建与生成

研究背景与意义

当前重建技术面临的挑战

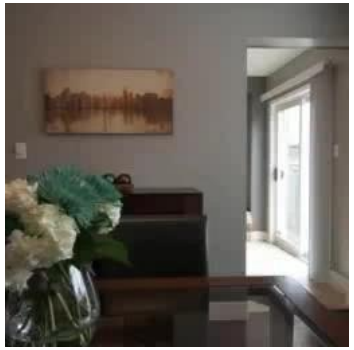
- 依赖多视角稠密采集
- 无法重建视角未覆盖区域

当前生成技术面临的挑战



- 精准可控性
- 时空一致性

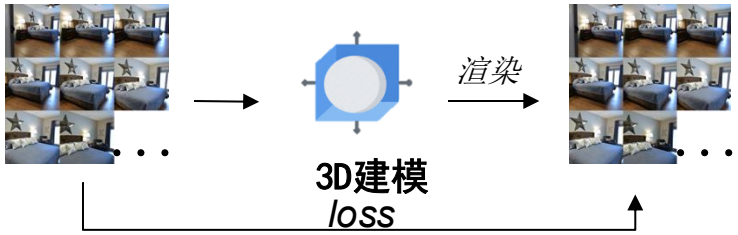
重建与生成结合



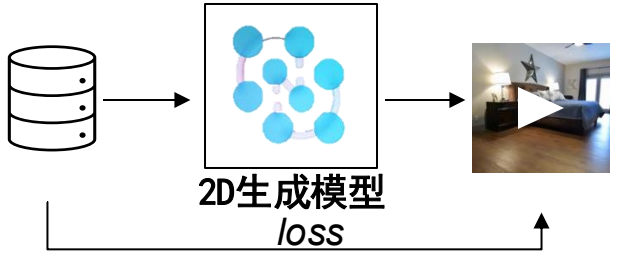
- 实现稀疏视角下完整的场景重建
- 实现精准可控且3D一致的长视频生成

重建与生成结合：空间智能生成模型的有效实现路径

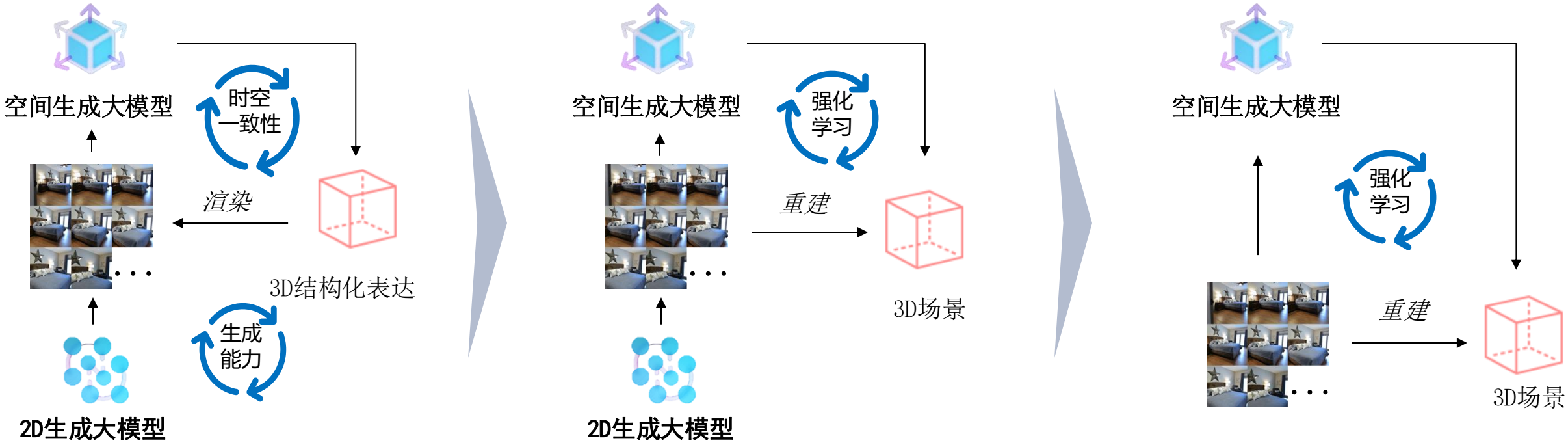
传统重建与渲染管线：产出效率低



当前视频生成大模型：缺乏时空一致性



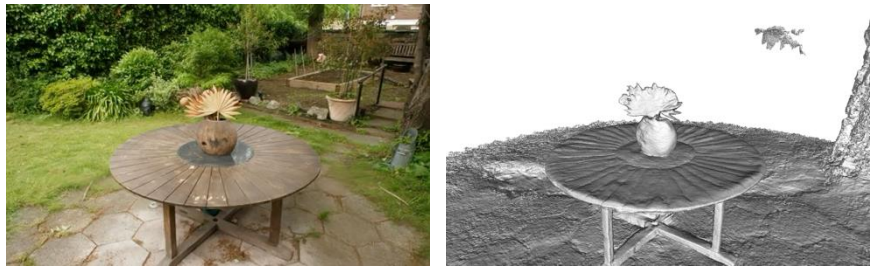
空间智能生成模型框架（分三个阶段）



阶段一工作：重建大模型与视频生成大模型的结合

- 极度稀疏视角下的三维重建与生成

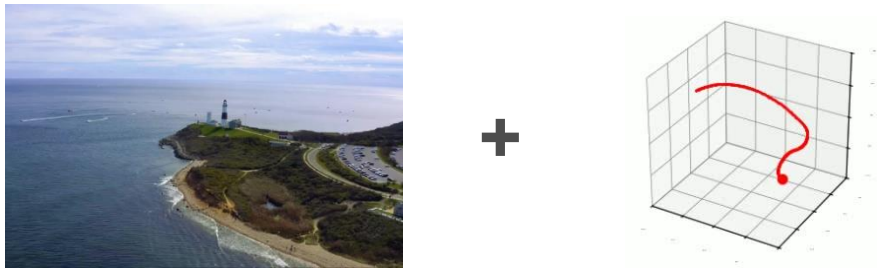
稀疏视角



3DGS重建　　表面重建

- 基于自回归框架的高度可控、几何一致的场景生成

单帧图像控制　　相机控制

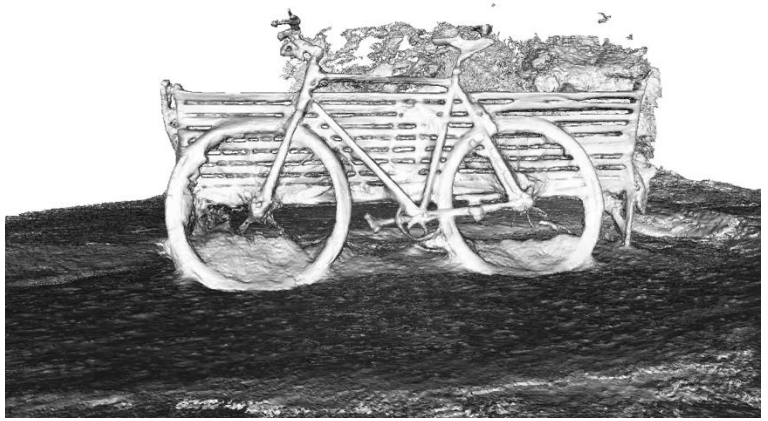


生成结果

Free360: Layered Gaussian Splatting for Unbounded 360-Degree View Synthesis from Extremely Sparse and Unposed Views. CVPR 2025.

StarGen: A Spatiotemporal Autoregression Framework with Video Diffusion Model for Scalable and Controllable Scene Generation. CVPR 2025.

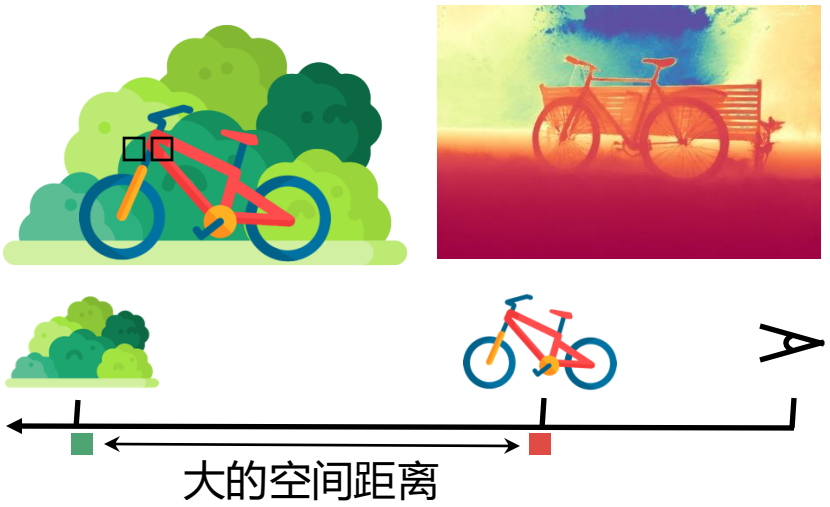
极度稀疏视角和未知位姿的360°无界场景层次化高斯溅射



无界场景中**极度稀疏** (3-4) 和未知位姿的图像

360° 全新视角生成和表面重建

极度稀疏视角和未知位姿的360°无界场景层次化高斯溅射



无界场景的空间歧义性

层次化高斯溅射表达

稀疏立体重建中的几何噪声

逐层的重建引导优化

视频生成模型中的多视角不一致性

重建与生成的迭代融合

极度稀疏视角和未知位姿的360°无界场景层次化高斯溅射

稀疏的输入视角

前景层图像和单目法向监督信号

前景层初始重建　　前景层三维高斯

稠密立体重建模型

逐层的重建引导优化

被优化的前景层重建

稠密的背景层重建

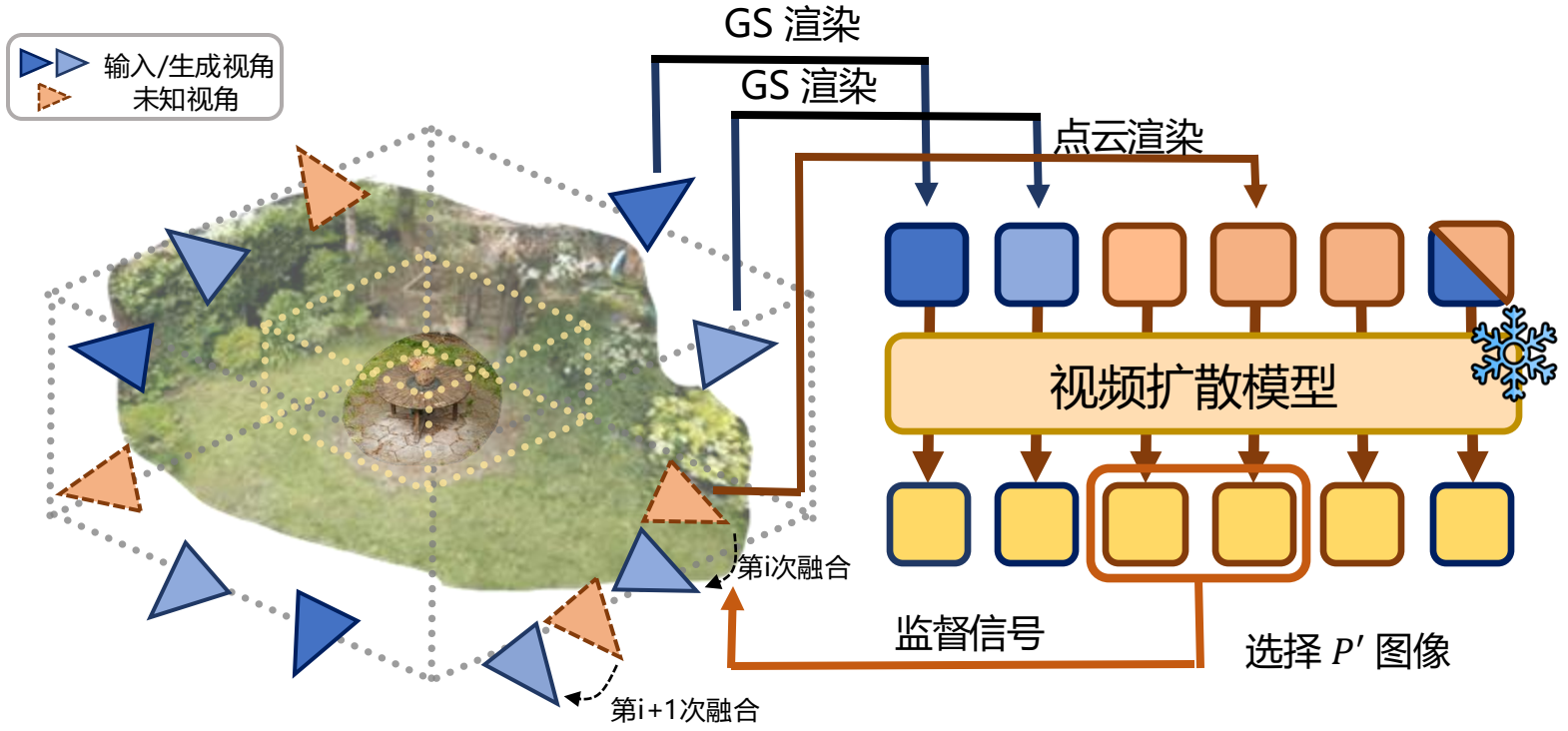
多层重建结果点云可视化

极度稀疏视角和未知位姿的360°无界场景层次化高斯溅射

被优化的前景层重建

稠密的背景层重建

多层重建结果点云可视化



分层高斯溅射渲染

重建与生成的迭代融合

实验结果

稀疏图像输入

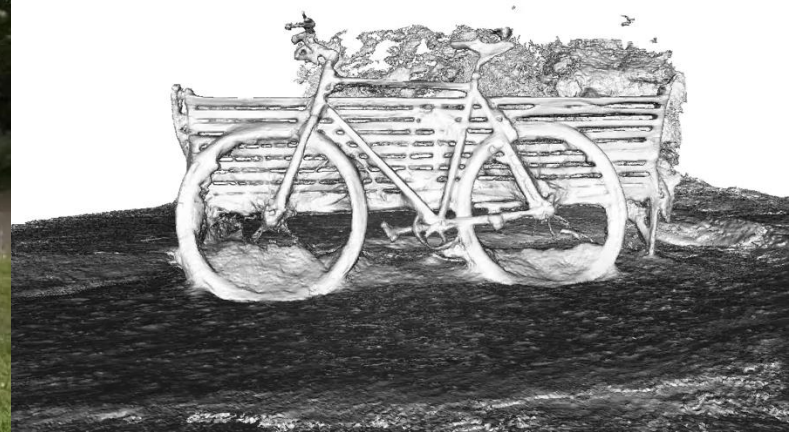
FSGS*

InstantSplat

ZeroNVS*

ViewCrafter

我们的渲染结果

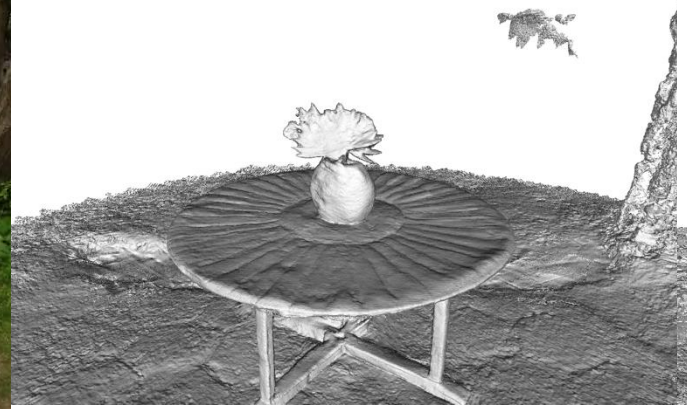


我们的表面重建结果

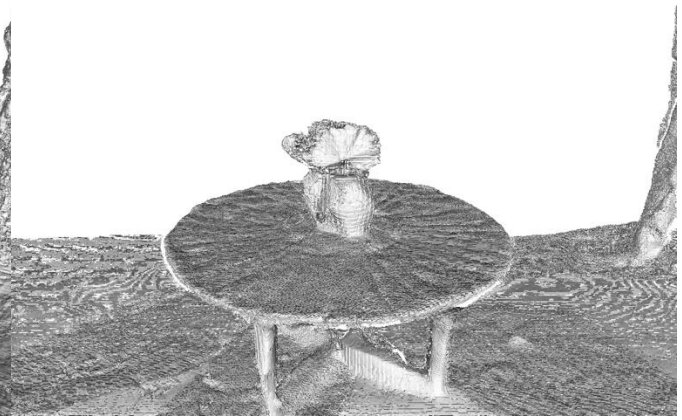
实验结果

稀疏图像输入

我们的渲染结果



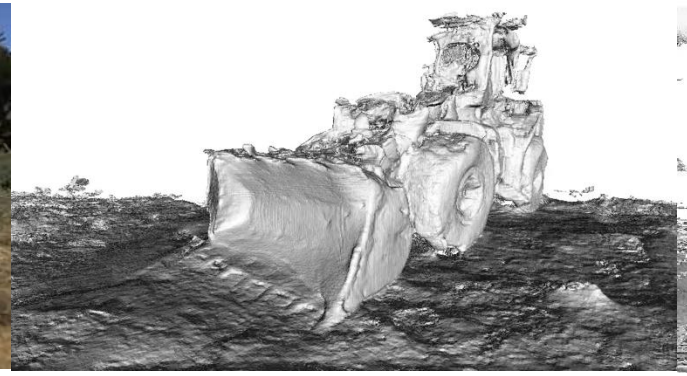
我们的表面重建结果



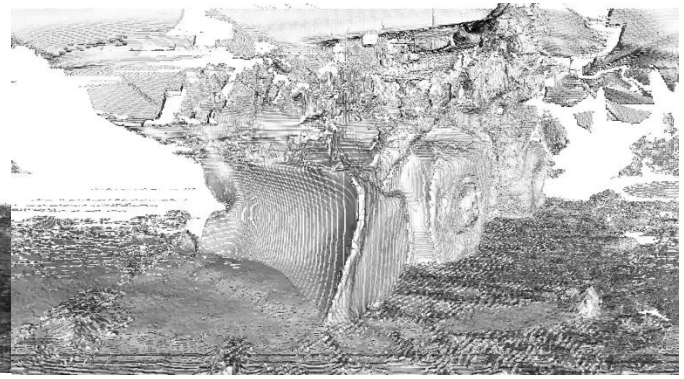
2DGS*

稀疏图像输入

我们的渲染结果

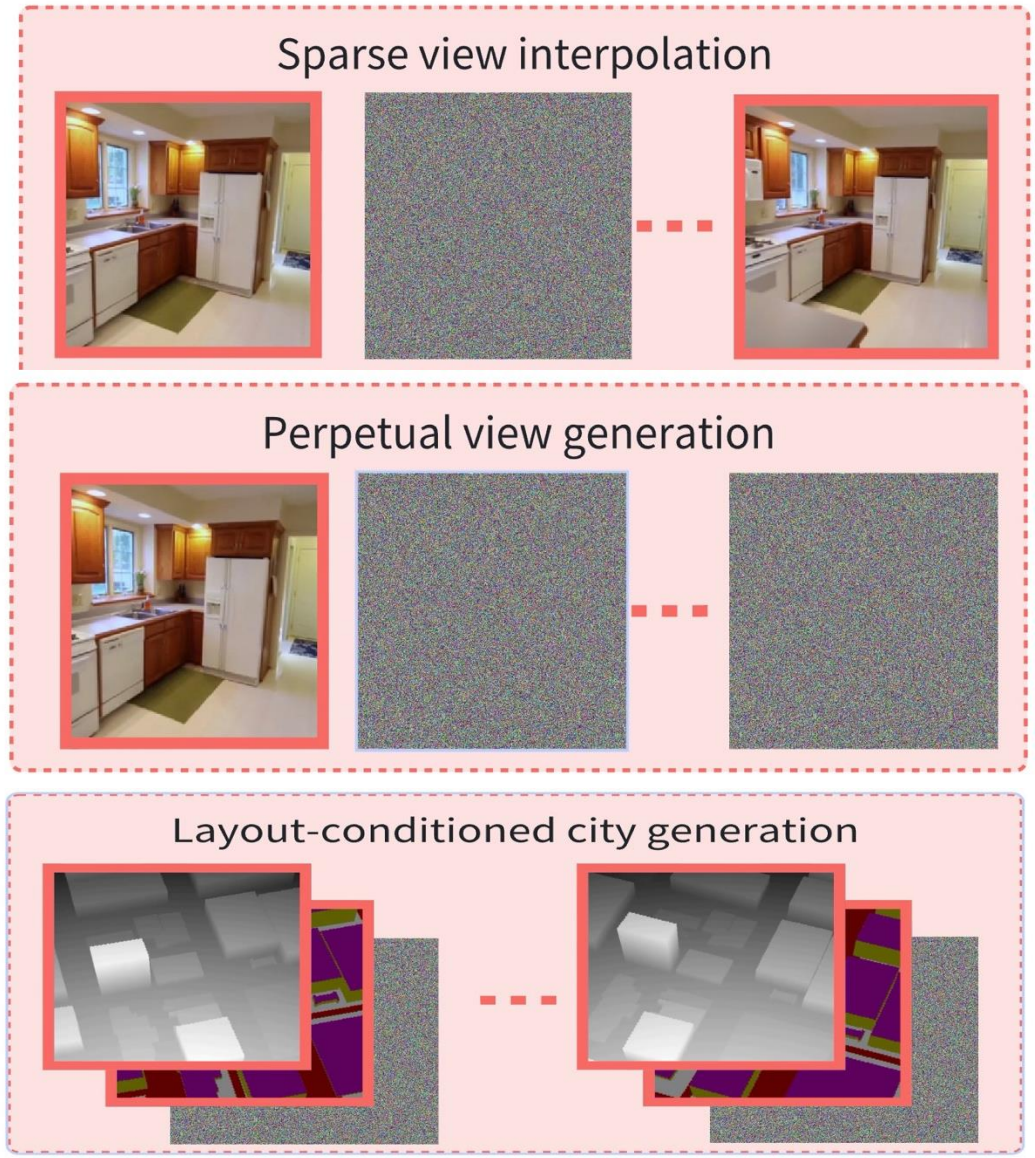


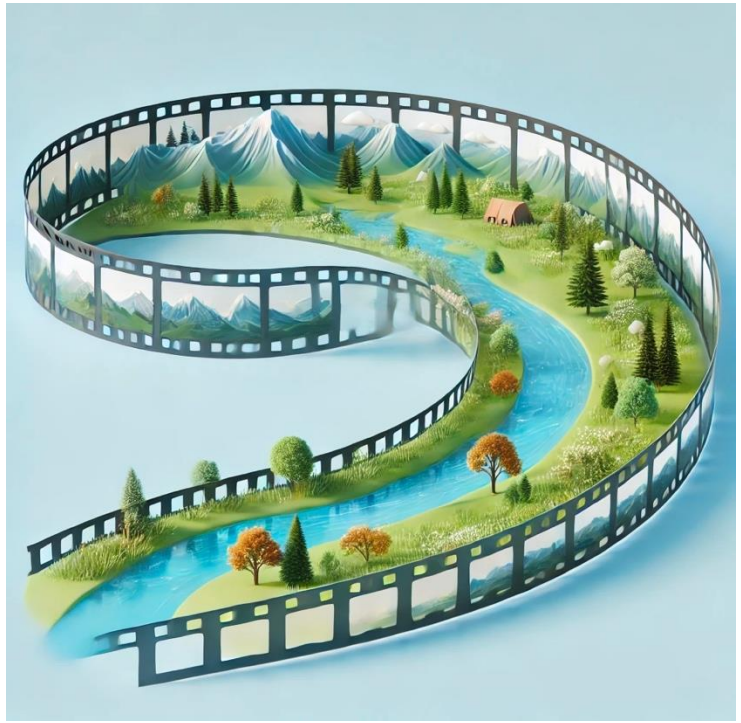
我们的表面重建结果



2DGS*

多样的场景生成任务的通用框架





多视一致与可控位姿的长视频生成方法

关键问题与挑战

重建模型本身的误差

↓

重建模型和生成模型结合

长距离视频间的多视一致

↓

多视一致的可控视频生成

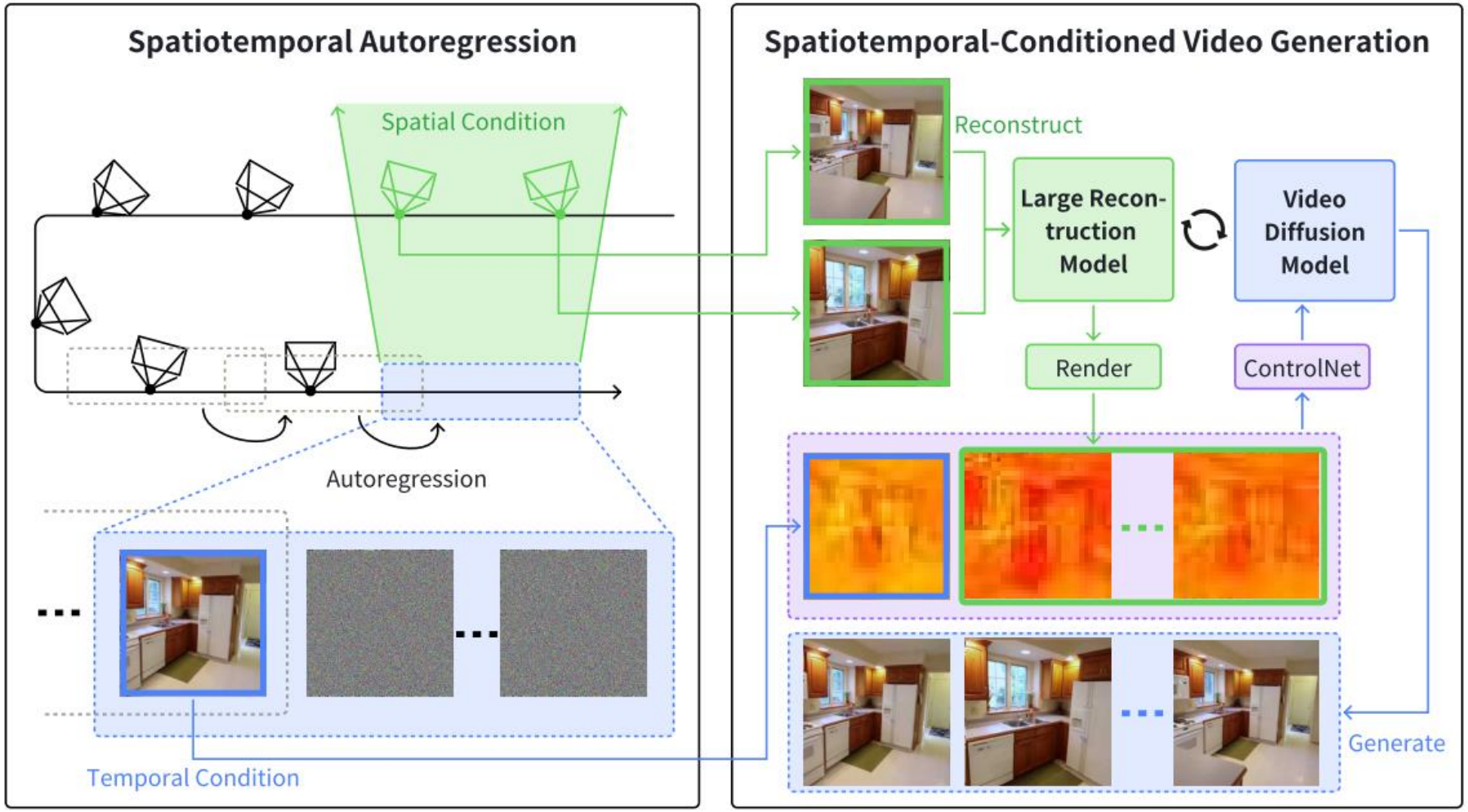
自回归视频的误差累计

↓

时空一致的自回归框架

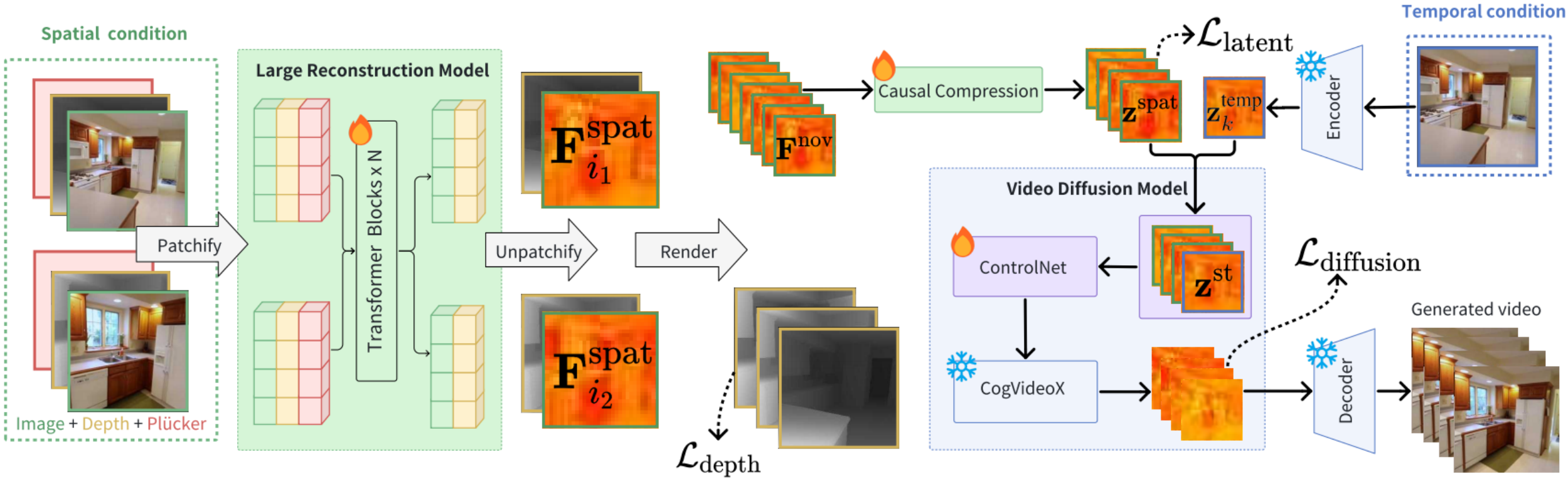
时空一致的自回归生成框架

以自回归形式生成长视频，结合空间邻近图像和时间图像，始终保证每段视频与场景保持一致，从而实现整个长视频的多视一致。



多视一致的可控视频生成

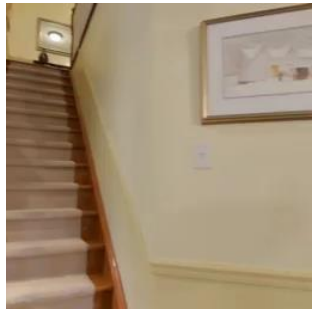
- 空间邻近图像通过重建模型生成特征和深度图，并渲染到生成视角下，并进一步压缩到VAE空间。
- 时间条件图像压缩到VAE空间后，替换对应的空间特征形成**时空条件特征。**
- 通过ControlNet，时空条件特征注入生成模型，保证每段视频的**时空一致性。**



实验结果

输入

首帧



尾帧

生成结果

Groundtruth

145 frames

StarGen (ours)

290 frames

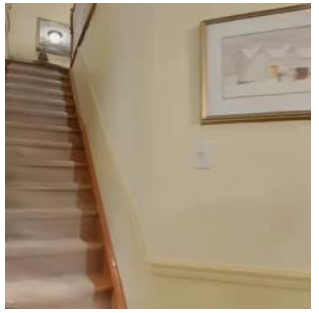
ViewCrafter

25 frames

MVSplat

145 frames

DepthSplat

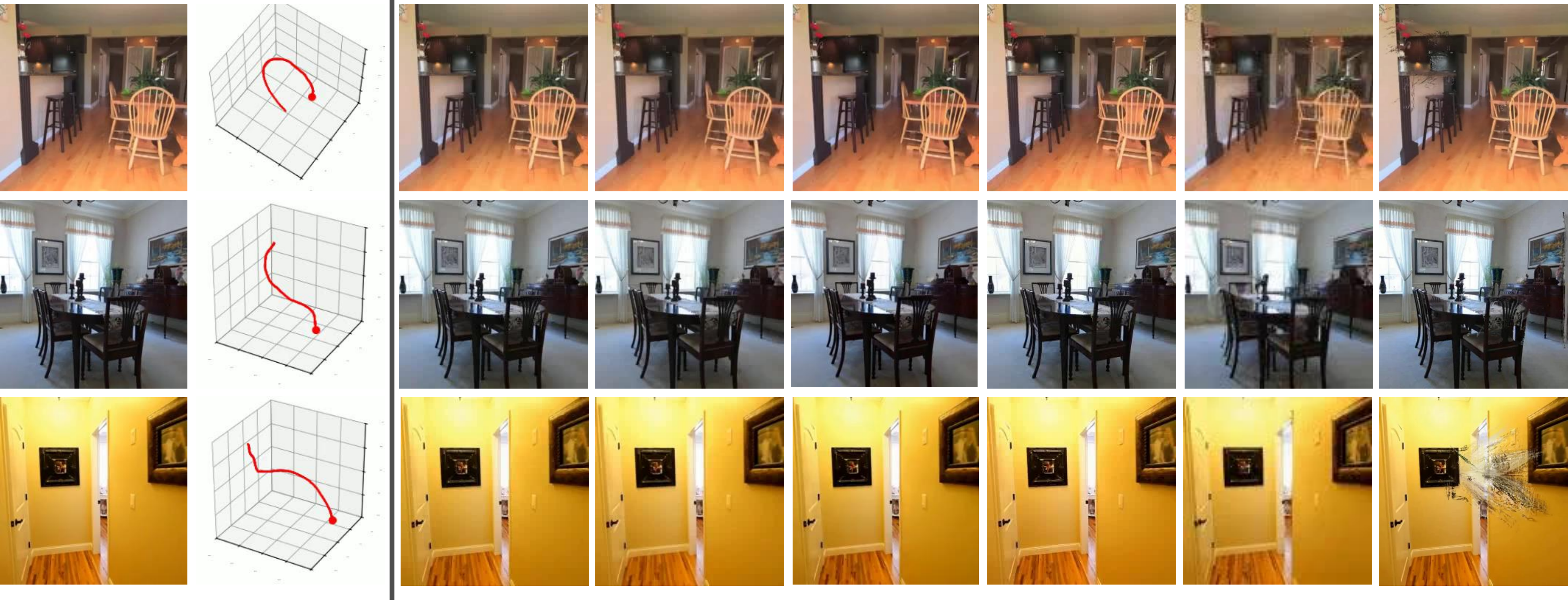


145 framess

实验结果

输入 | 生成结果

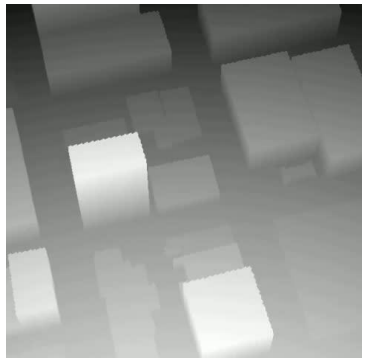
首帧 | 位姿控制 | GT | StarGen (ours) | ViewCrafter | MotionCtrl | InfNat0 | LucidDreamer

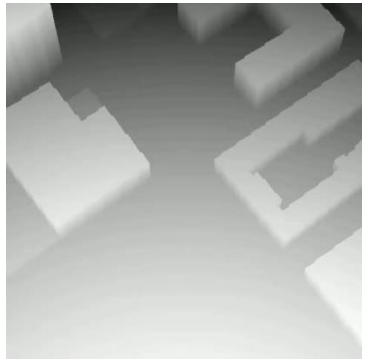


实验结果

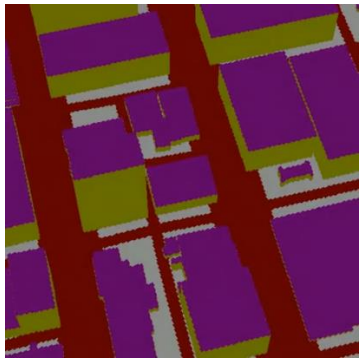
输入

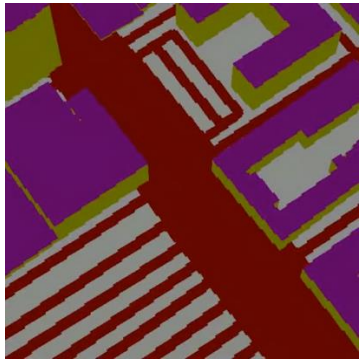
深度条件

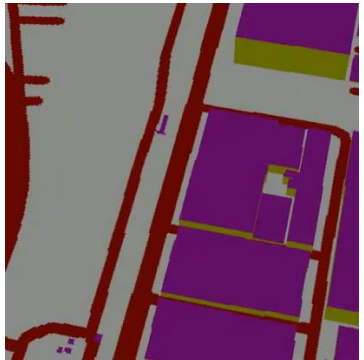




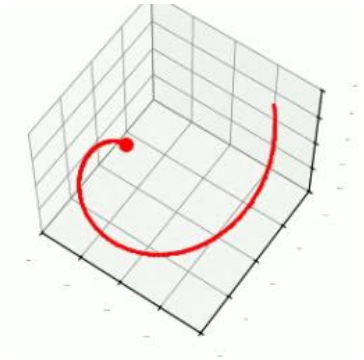
语义条件

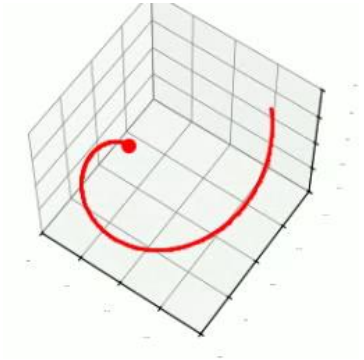


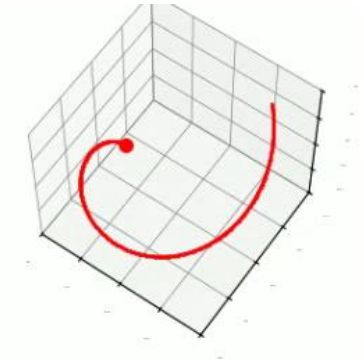




位姿控制

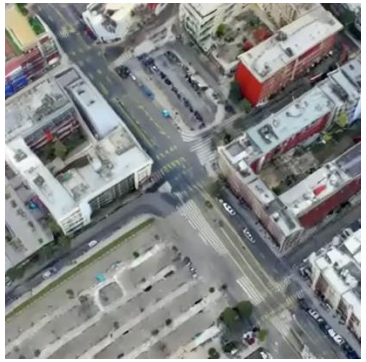


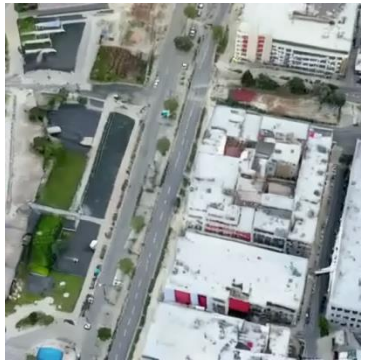




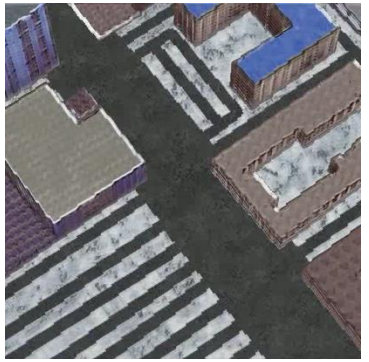
生成结果

StarGen (ours)





CityDreamer

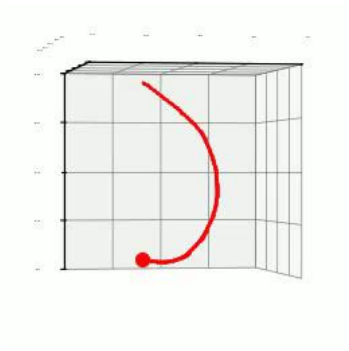


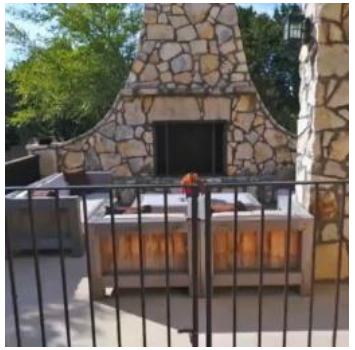
实验结果

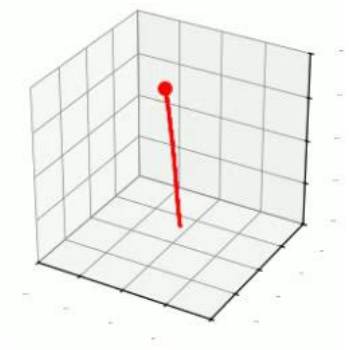
输入

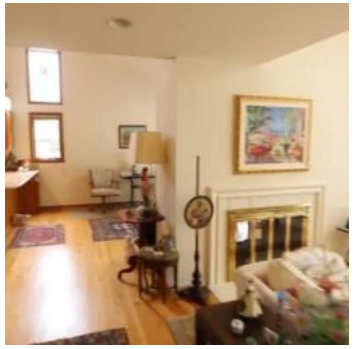
首帧

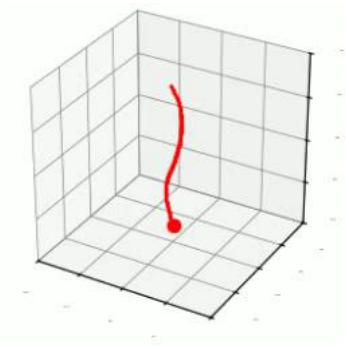
位姿控制





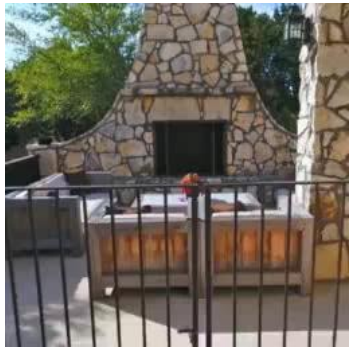


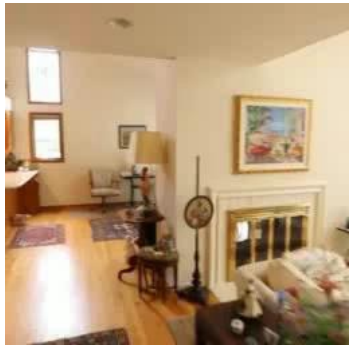




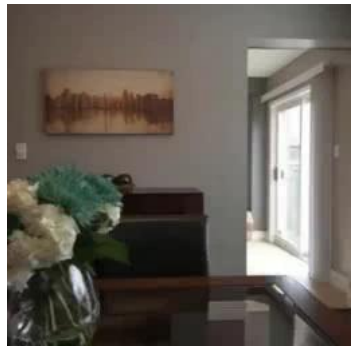
生成结果

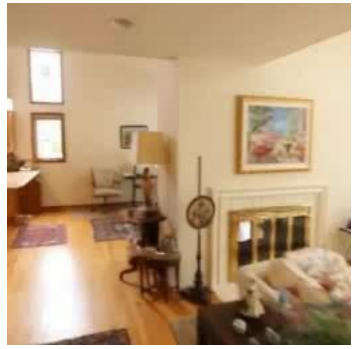
Groundtruth



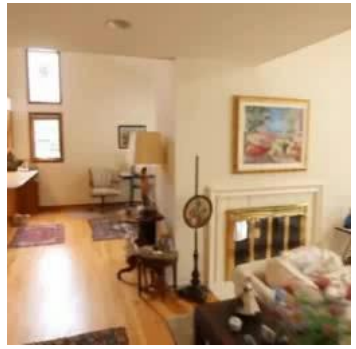


Ours

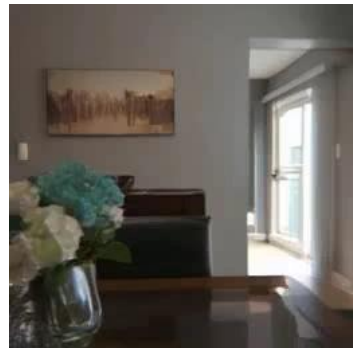


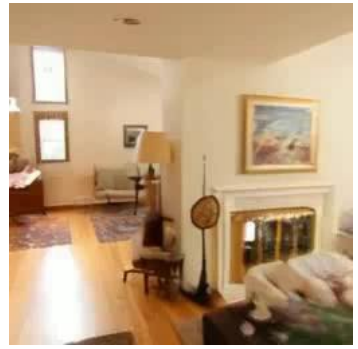


w/o spatial cond.

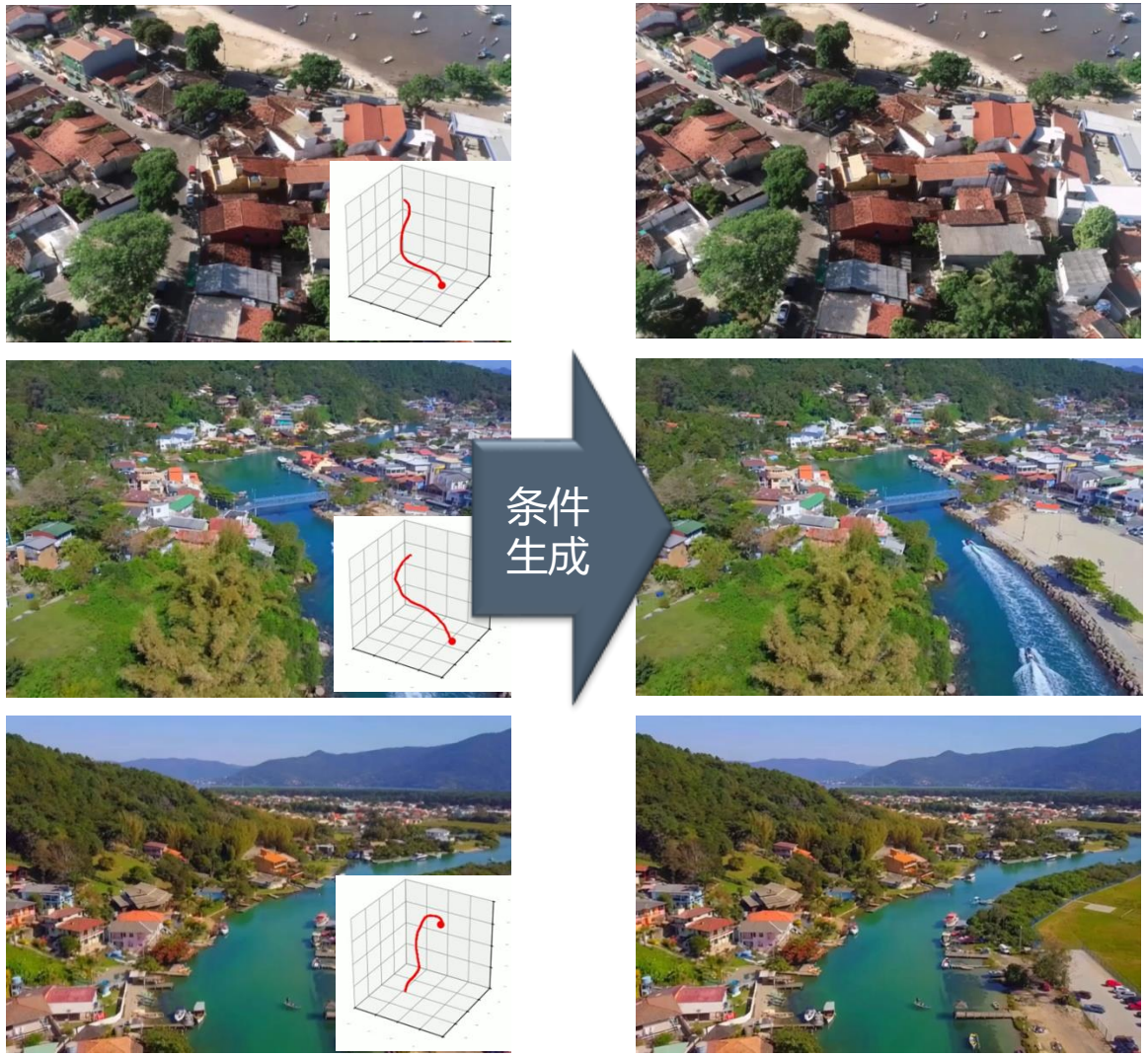


w/o temporal cond.



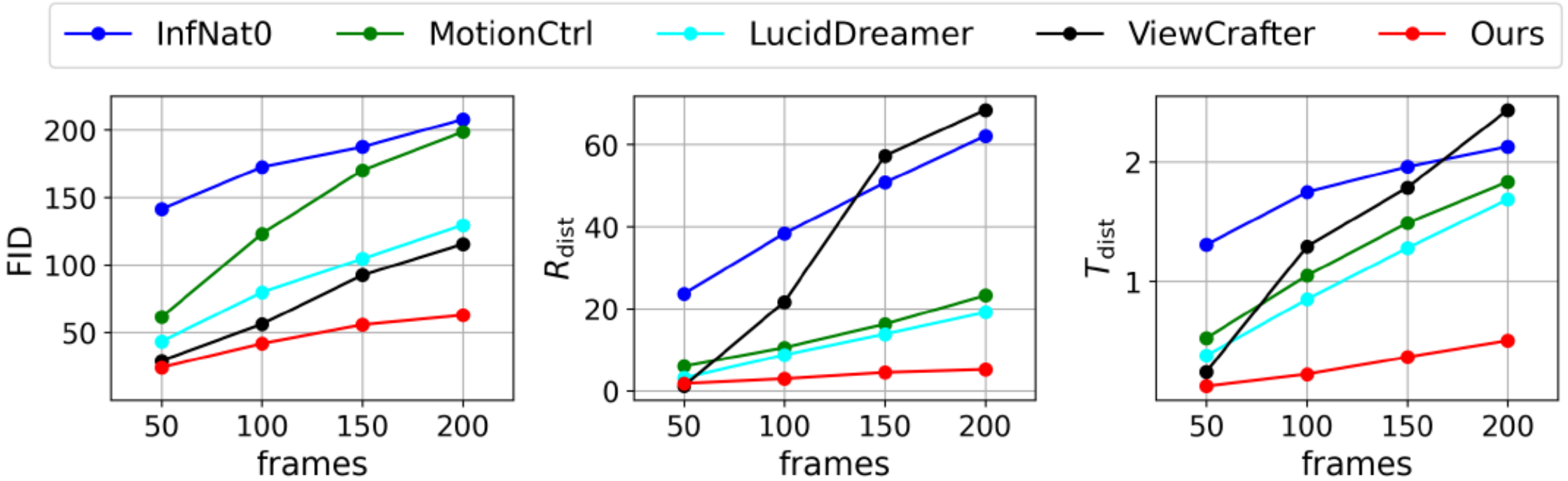


更多单视图场景生成例子



定量比较

- 得益于3D重建大模型和视频生成大模型的有机结合，随着生成视频的帧数增加，我们的方法在**生成质量**和**运动控制精度**上均优于SOTA

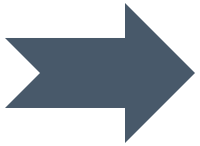


注：FID度量生成质量，R_{dist}/T_{dist}度量运动控制的旋转/平移误差

3D场景生成

输入一张图像

StarGen+3DGS



总结与展望

- 随着大模型的发展，**重建**与**生成**任务已开始融合
  - Free360：利用**生成大模型**补全不可见区域，实现**极度稀疏视角的三维重建**；
  - StarGen：利用**重建大模型**重建并渲染作为控制条件，实现**时空一致的长视频生成**
- 未来发展趋势
  - 重建与生成进一步融合，最终实现同时具备重建与生成能力的**原生三维**大模型
  - 如何在**移动端实时**，芯片化和端云协同将是一个趋势

谢 谢！

http://www.cad.zju.edu.cn/home/gfzhang/
https://github.com/zju3dv

Email: zhangguofeng@zju.edu.cn