

ViewCrafter: 驯服视频扩散模型以实现高保真新视角合成

ViewCrafter: Taming Video Diffusion Models for High-fidelity Novel View Synthesis

Wangbo Yu*, Jinbo Xing*, Li Yuan*, Wenbo Hu†, Xiaoyu Li, Zhipeng Huang,
Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian†, *Fellow, IEEE*

摘要—Despite recent advancements in neural 3D reconstruction, the dependence on dense multi-view captures restricts their broader applicability. In this work, we propose **ViewCrafter**, a novel method for synthesizing high-fidelity novel views of generic scenes from single or sparse images with the prior of video diffusion model. Our method takes advantage of the powerful generation capabilities of video diffusion model and the coarse 3D clues offered by point-based representation to generate high-quality video frames with precise camera pose control. To further enlarge the generation range of novel views, we tailored an iterative view synthesis strategy together with a camera trajectory planning algorithm to progressively extend the 3D clues and the areas covered by the novel views. With ViewCrafter, we can facilitate various applications, such as immersive experiences with real-time rendering by efficiently optimizing a 3D-GS representation using the reconstructed 3D points and the generated novel views, and scene-level text-to-3D generation for more imaginative content creation. Extensive experiments on diverse datasets demonstrate the strong generalization capability and superior performance of our method in synthesizing high-fidelity and consistent novel views. Our project webpage and code are available at <https://drexubery.github.io/ViewCrafter/>.

神经三维重建近来虽已取得长足进展，但它对稠密多视角采集的依赖限制了应用范围。本文提出 **ViewCrafter**，一种借助视频扩散模型先验、从单张或稀疏图像合成通用场景高保真新视角的方法。该方法一方面发挥视频扩散模型强大的生成能力，另一方面利用基于点的表示所提供的粗略三维线索，从而在精确控制相机位姿的前提下生成高质量视频帧。为进一步扩大新视角的生成范围，我们量身设计了一套迭代式视角合成策略，并辅以相机轨迹规划算法，逐步扩展三维线索与新视角覆盖的区域。借助 ViewCrafter，我们能够支撑多种应用：例如利用重建出的三维点与生成的新视角高效优化 3D 高斯泼溅（3D-GS）表示，实现实时渲染的沉浸式体验；又如面向更富想象力内容创作的场景级文本到 3D 生成。在多样数据集上的大量实验表明，本方法在合成高保真且一致的新视角上具有很强的泛化能力与优越性能。项目主页与代码见 <https://drexubery.github.io/ViewCrafter/>。

Index Terms—新视角合成，视频扩散模型，三维场景生成。



1 引言 (Introduction)

Novel view synthesis plays a crucial role in computer vision and graphics for creating immersive experiences in games, mixed reality, and visual effects. Despite the significant success of 3D neural reconstruction techniques such as NeRF [1] and 3D-GS [2], their dependence on dense multi-view observations

restricts their broader applicability in situations where only limited views are available.

视角合成在计算机视觉与图形学中扮演着关键角色，可为新游戏、混合现实和视觉特效创造沉浸式体验。尽管 NeRF [1]、3D-GS [2] 等三维神经重建技术已取得显著成功，但它们依赖稠密的多视角观测，在仅有少量视角可用的场景中适用性受限。

A more desirable problem scenario in practice involves synthesizing novel views of generic scenes from sparse observations or even a single image. This task is considerably challenging as it necessitates a comprehensive understanding of the 3D world, including structures, appearance, semantics,

- *: Equal contribution. †: Corresponding authors. Wangbo Yu, Li Yuan, Zhipeng Huang, Yonghong Tian are with Peking University and Peng Cheng Laboratory. Jinbo Xing is with The Chinese University of Hong Kong. Wenbo Hu, Xiaoyu Li are with Tencent AI Lab. Ying Shan is with ARC Lab, Tencent PCG and Tencent AI Lab. Xiangjun Gao is with Hong Kong University of Science and Technology. Tien-Tsin Wong is with Monash University.
- 本文为 AI 辅助翻译版本，仅供学习交流；引用请以英文原文为准：arXiv:2409.02048。

and occlusions. Early researches [3], [4], [5], [6], [7], [8], [9], [10] focused on training regression-based models to synthesize novel views from sparse or single input. However, due to their limited representation capabilities, these methods are mostly category-specific and only handle certain domains such as indoor scenes. Recent advancements in powerful diffusion models have made zero-shot novel view synthesis [11], [12], [13] from single view approachable. Nevertheless, these methods are either restricted to handling object-level images or lack precise control of the camera pose due to their dependency on high-level pose prompts to guide the view synthesis process. Some works [14], [15] also attempt to synthesize novel views from a single image using depth-based warping and diffusion-based inpainting. Yet, these methods often produce inconsistent content in occlusion regions due to the per-view inpainting mechanism.

实践中更理想的问题设定，是从稀疏观测乃至单张图像出发，合成通用场景的新视角。这一任务相当困难，因为它要求模型全面理解三维世界，包括结构、外观、语义以及遮挡关系。早期研究 [3], [4], [5], [6], [7], [8], [9], [10] 着眼于训练基于回归的模型，从稀疏或单张输入合成新视角。然而受限于表示能力，这类方法大多针对特定类别，只能处理室内场景等有限领域。近来强大的扩散模型不断进步，使从单一视角进行零样本新视角合成 [11], [12], [13] 成为可能。但这些方法要么只能处理物体级图像，要么因依赖高层位姿提示来引导视角合成而无法精确控制相机位姿。也有一些工作 [14], [15] 尝试用基于深度的图像变形加扩散式修补，从单张图像合成新视角；然而受制于逐视角修补的机制，它们在遮挡区域往往生成前后不一致的内容。

In this work, we focus on high-fidelity novel view synthesis of *generic scenes* from single or sparse images, maintaining *precise control of the camera pose* and *consistency* of the generated novel views. To achieve this, we investigate leveraging the generative capabilities of video diffusion models alongside the explicit 3D information provided by point cloud representations. On one hand, video diffusion models [16], [17], [18], trained on web-scale video datasets, develop a reasonable understanding of the world, which facilitates the generation of plausible video content from a single image or text prompt. However, they lack the underlying 3D information of the scene and fall short in achieving precise camera view control. On the other hand, recent dense stereo methods [19], [20] have made fast point cloud reconstruction from single or sparse images accessible. Point cloud representation provides valuable coarse 3D scene information, enabling precise pose control for free-view rendering. Yet, due to its poor representation capability and the limited 3D cues offered by extremely sparse reference images, it suffers from occlusions, missing areas, and geometry distortion, hindering its utility in novel view synthesis. With these in mind, we propose integrating the generative power of

video diffusion models with the coarse 3D prior provided by point-based representation, aiming to facilitate higher-fidelity novel view synthesis of generic scenes.

本文聚焦于从单张或稀疏图像出发、对通用场景进行高保真新视角合成，同时保持对相机位姿的精确控制以及所生成新视角的一致性。为此，我们探索将视频扩散模型的生成能力与点云表示所提供的显式三维信息结合起来。一方面，在网络规模视频数据上训练的视频扩散模型 [16], [17], [18] 形成了对世界较为合理的理解，能够从单张图像或文本提示生成合理的视频内容；但它们缺乏场景底层的三维信息，难以实现精确的相机视角控制。另一方面，近来的稠密立体方法 [19], [20] 让从单张或稀疏图像快速重建点云成为可能。点云表示提供了宝贵的粗略三维场景信息，可支持自由视角渲染中的精确位姿控制；但受限于其表示能力不足，加之极稀疏参考图像所能提供的三维线索有限，点云会出现遮挡、缺失区域和几何畸变，妨碍其在新视角合成中的应用。有鉴于此，我们提出把视频扩散模型的生成能力与基于点的表示所提供的粗略三维先验融合起来，以实现通用场景更高保真的新视角合成。

Our method, **ViewCrafter**, accomplishes novel view synthesis by a *point-conditioned video diffusion model* that generates high-fidelity and consistent videos under a novel view trajectory, conditioned on corresponding frames rendered from point cloud reconstructed from single or sparse images. Leveraging the explicit 3D information from the point cloud and the generative capabilities of video diffusion models, our method enables precise control of 6 DoF camera poses and generates high-fidelity, consistent novel views. Furthermore, video diffusion models face challenges in generating long videos because of unacceptably increased memory and computation costs. To tackle this challenge, we propose an *iterative view synthesis* strategy along with a *content-adaptive camera trajectory planning* algorithm to progressively extend the reconstructed point cloud and the areas covered by the novel views. Starting from the initial point cloud derived from the input image(s), we first employ the camera trajectory planning algorithm to predict a content-adaptive camera pose sequence from the current point cloud to effectively reveal occlusions. Next, we render the point cloud according to the predicted pose sequence and synthesize novel views by ViewCrafter with the conditions of the rendered point cloud. Subsequently, the point cloud is updated from the synthesized novel views to extend the global point cloud representation. Through iteratively conducting these steps, we can ultimately obtain high-fidelity novel views that cover a large view range and an extended point cloud.

我们的方法 **ViewCrafter** 通过一个点云条件视频扩散模型来完成新视角合成：以从单张或稀疏图像重建的点云所渲染的对应帧为条件，沿新视角轨迹生成高保真、一致的视

频。借助点云提供的显式三维信息与视频扩散模型的生成能力，本方法能够精确控制 6 自由度相机位姿，并生成高保真、一致的新视角。此外，由于内存和计算开销会不可接受地激增，视频扩散模型在生成视频时面临困难。为应对这一挑战，我们提出一套迭代式视角合成策略，并配以内容自适应的相机轨迹规划算法，逐步扩展重建出的点云以及新视角覆盖的区域。从输入图像得到的初始点云出发，我们先用相机轨迹规划算法，基于当前点云预测一条内容自适应的相机位姿序列，以有效揭示遮挡区域；随后依照预测的位姿序列渲染点云，并以渲染结果为条件用 ViewCrafter 合成新视角；接着再用合成出的新视角更新点云，扩展全局点云表示。反复执行上述步骤，最终即可得到覆盖大视角范围的高保真新视角以及一份扩展后的点云。

In addition to novel view synthesis, we explore several applications of our method. For instance, we can efficiently optimize a 3D-GS representation based on the constructed point cloud and the synthesized novel views within minutes, enabling real-time rendering for immersive experiences. Furthermore, our method shows the potential to adapt to scene-level text-to-3D generation, which can foster more imaginative artistic creations.

除新视角合成外，我们还探索了本方法的若干应用。例如，基于构建好的点云和合成的新视角，我们可以在数分钟内高效优化一个 3D-GS 表示，从而实现面向沉浸式体验的实时渲染。此外，本方法还展现出适配场景级文本到 3D 生成的潜力，可催生更富想象力的艺术创作。

We extensively evaluate our method for zero-shot novel view synthesis and sparse view 3D-GS reconstruction on various datasets, including Tanks-and-Temples [21], RealEstate10K [7], and CO3D [22]. For zero-shot novel view synthesis, our method outperforms the baselines in both image quality and pose accuracy metrics. This demonstrates its superior ability to synthesize high-fidelity novel views and achieve precise pose control. In 3D-GS reconstruction, our approach consistently surpasses previous state-of-the-art. This further validates its effectiveness in scene reconstruction from sparse views.

我们在 Tanks-and-Temples [21]、RealEstate10K [7] 和 CO3D [22] 等多个数据集上，对本方法在零样本新视角合成与稀疏视角 3D-GS 重建两方面进行了充分评测。在零样本新视角合成上，本方法在图像质量与位姿精度两类指标上均优于基线，表明其在合成高保真新视角、实现精确位姿控制上更胜一筹。在 3D-GS 重建上，本方法始终超越以往最先进方法，进一步印证了它在稀疏视角场景重建中的有效性。

Our contributions can be summarized as follows:

本文的贡献可归纳如下：

- We propose **ViewCrafter**, a novel view synthesis framework tailored for synthesizing high-fidelity novel view sequences of generic scenes from single or sparse images while maintaining precise control of camera poses.

我们提出 **ViewCrafter**，一个专为从单张或稀疏图像合成通用场景高保真新视角序列、并能精确控制相机位姿而设计的新视角合成框架。

- We present an iterative view synthesis strategy in conjunction with a content-adaptive camera trajectory planning algorithm to progressively expand the covered areas of novel views and the reconstructed point cloud, enabling long-range and large-area novel view synthesis. We propose a set of iterative view synthesis strategies, and combine content-adaptive camera trajectory planning algorithms, to progressively expand the covered areas of novel views and the reconstructed point cloud,从而实现长距离、大范围的新视角合成。
- Our method achieves superior performance on various challenging datasets in terms of both the quality of synthesized novel views and the accuracy of camera pose control. It facilitates various applications beyond novel view synthesis, such as real-time rendering for immersive experiences by efficiently optimizing a 3D-GS representation from our results, and scene-level text-to-3D generation for more imaginative artistic creations.

本方法在多个具有挑战性的数据集上，无论是合成新视角的质量还是相机位姿控制的精度都取得了优越表现。它还能支撑新视角合成之外的多种应用，例如从我们的结果高效优化 3D-GS 表示以实现面向沉浸式体验的实时渲染，以及面向更富想象力艺术创作的场景级文本到 3D 生成。

2 相关工作 (Related Work)

2.1 基于回归的新视角合成 (Regression-based Novel View Synthesis)

Regression-based methods aim to train a feed-forward model to generate novel views from sparse or single image inputs. This is often achieved using CNN/Transformer-based [23] architectures to establish a 3D representation of the input image(s). For instance, several works [24], [25] have applied this idea to specific modalities, such as human faces, by generating tri-plane representations for novel view synthesis. LRM [26] extends this strategy to generic objects, while other methods like [7], [8], [9] adopt the multi-plane representation, and PixelNeRF [10] employs NeRF [1] as 3D representation for novel view synthesis. Inspired by the success of 3D-GS [2], recent approaches such as PixelSplat [27] and MVSplat [28] explore training regression-based models to produce 3D Gaussian representations for real-time rendering capabilities. Additionally, some methods like

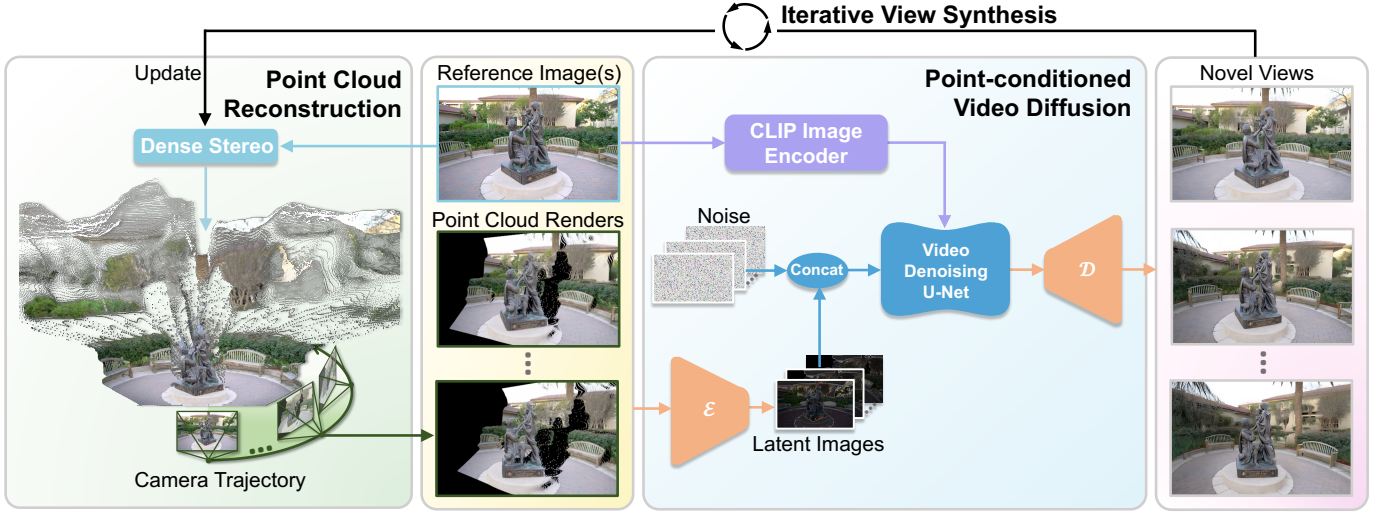


图 1. **ViewCrafter 方法总览**。给定单张参考图像或稀疏图像集，我们首先用稠密立体模型构建其点云表示，从而能够精确移动相机进行自由视角渲染。随后，为解决点云渲染结果中出现的大面积缺失区域、几何畸变和点云伪影，我们训练一个点云条件视频扩散模型作为增强渲染器，基于粗略的点云渲染生成高保真、一致的新视角。为实现长距离新视角合成，我们采用迭代式视角合成策略，反复移动相机、生成新视角并更新点云，从而实现更完整的点云重建，并惠及 3D-GS 优化等下游任务。**Overview of our ViewCrafter.** Given a single reference image or sparse image sets, we first build its point cloud representation using a dense stereo model, which enables accurately moving cameras for free-view rendering. Subsequently, to address the large missing regions, geometric distortions, and point cloud artifacts exhibited in the point cloud render results, we train a point-conditioned video diffusion model to serve as an enhanced renderer, facilitating the generation of high-fidelity and consistent novel views based on the coarse point cloud renders. To achieve long-range novel view synthesis, we adopt an iterative view synthesis strategy that involves iteratively moving cameras, generating novel views, and updating the point cloud, which enables a more complete point cloud reconstruction and benefits downstream tasks such as 3D-GS optimization.

[3], [4], [5], [6] combine monocular depth estimation and image inpainting modules in a unified framework for novel view synthesis. However, these methods are limited to category-specific domains, such as objects and indoor scenes, and are prone to artifacts due to their limited model representation capabilities. In contrast, our method can synthesize high-fidelity novel views of generic scenes.

基于回归的方法旨在训练一个前馈模型，从稀疏或单张图像输入生成新视角。这通常借助基于 CNN/Transformer [23] 的架构，为输入图像建立三维表示来实现。例如，一些工作 [24], [25] 把这一思路用于人脸等特定模态，通过生成三平面表示来做新视角合成。LRM [26] 将该策略推广到通用物体；[7], [8], [9] 等方法采用多平面表示；PixelNeRF [10] 则以 NeRF [1] 作为三维表示来合成新视角。受 3D-GS [2] 成功的启发，PixelSplat [27]、MVSplat [28] 等近期方法探索训练基于回归的模型来输出三维高斯表示，以获得实时渲染能力。此外，[3], [4], [5], [6] 等方法将单目深度估计与图像修补模块统一到一个框架中来合成新视角。然而，这些方法局限于特定类别的领域（如物体和室内场景），且因模型表示能力有限而易产生伪影。相比之下，本方法能够合成通用场景的高保真新视角。

2.2 基于扩散模型的新视角合成 (Diffusion-based Novel View Synthesis)

The rapid advancement of diffusion models [29], [30], [31] have demonstrated exceptional proficiency in synthesizing high-quality images and shows the potential to be adapted in synthesizing novel views from single or sparse inputs. While some optimization-based approaches [32], [33], [34] directly train a 3D representation under the supervision of text-to-image (T2I) diffusion models [31], they require scene-specific optimization, which compromising their generalization capabilities. To address this, GenVNS [35] proposes a generalized novel view synthesis framework by training a 3D feature-conditioned diffusion model on a large-scale multiview dataset [22]. Similarly, Zero-1-to-3 [11] develops camera pose-conditioned diffusion models trained on synthetic datasets [36], [37], enabling novel view synthesis from more diverse inputs. However, these models are either category-specific [35], [38] or limited to handling toy-level objects with simple backgrounds. Recently, ZeroNVS [12] improves the generation capability of Zero-1-to-3 [11] by training it on a mixed dataset containing both synthetic [36] and real data [7], [39], [40], enabling zero-shot novel view synthesis of generic scenes from a single input image. Nonetheless, it still struggles to synthesize consistent novel views and lacks precise pose control, as it treats camera pose conditions as high-

level text embeddings. Reconfusion [41] proposes a PixelNeRF [10] feature-conditioned diffusion model to achieve relatively accurate pose control in novel view synthesis. However, it fails to synthesize consistent novel views due to its inability to model the correlations among sampled views. Additionally, it requires multiple images as input and cannot process a single image. Several works [14], [15], [42] utilize depth-based warping to synthesize novel views and employ a pre-trained T2I diffusion model [43] to refine the warped images. However, the novel views generated by these methods often suffer from artifacts and unrealistic contexts in the inpainted regions, which limits their applicability.

扩散模型 [29], [30], [31] 的迅猛发展在合成高质量图像上展现出卓越能力, 也显示出可用于从单张或稀疏输入合成新视角的潜力。一些基于优化的方法 [32], [33], [34] 直接在文本到图像 (T2I) 扩散模型 [31] 的监督下训练三维表示, 但它们需要针对具体场景进行优化, 损害了泛化能力。为了解决这一问题, GeNVS [35] 在大规模多视角数据集 [22] 上训练一个以三维特征为条件的扩散模型, 提出了一个可泛化的新视角合成框架。类似地, Zero-1-to-3 [11] 在合成数据集 [36], [37] 上训练以相机位姿为条件的扩散模型, 能够从更多样的输入合成新视角。然而这些模型要么针对特定类别 [35], [38], 要么只能处理背景简单的玩具级物体。近来, ZeroNVS [12] 在同时包含合成数据 [36] 与真实数据 [7], [39], [40] 的混合数据集上训练, 提升了 Zero-1-to-3 [11] 的生成能力, 实现了从单张输入图像对通用场景的零样本新视角合成。不过, 由于它把相机位姿条件当作高层文本嵌入处理, 仍难以合成一致的新视角, 也缺乏精确的位姿控制。Reconfusion [41] 提出一个以 PixelNeRF [10] 特征为条件的扩散模型, 在新视角合成中实现了相对精确的位姿控制; 但由于无法建模采样视角之间的关联, 它未能合成一致的新视角, 而且需要多张图像作为输入, 无法处理单张图像。还有一些工作 [14], [15], [42] 利用基于深度的图像变形来合成新视角, 并用预训练的 T2I 扩散模型 [43] 来修饰变形后的图像; 然而这些方法生成的新视角在修补区域常出现伪影和不真实的内容, 限制了其适用性。

2.3 条件式视频扩散模型 (Conditional Video Diffusion Models)

As generative models that produce content from diverse input modalities evolve rapidly, enhancing user control over generation has garnered significant interest. ControlNet [44], T2I-adapter [45], and GLIGEN [46] pioneered the introduction of condition signals for T2I generation. Similar strategies have also been employed in video generation, enabling controls like RGB images [17], [18], [47], depth [48], [49], trajectory [50], [51], and semantic maps [52]. However, camera motion control has received comparatively less attention. AnimateDiff [53]

and SVD [17] investigate class-conditioned video generation, grouping camera movements and utilizing LoRA [54] modules to create specific camera motions. MotionCtrl [13] improves control by using camera extrinsic matrices as conditioning signals. Although effective for simple trajectories, their dependence on 1D numeric values leads to imprecise control in complex real-world situations. MultiDiff [55] leverage depth-based warping to produce warped images, and condition the video diffusion model on the warped images to provide explicit 3D prior. Nevertheless, it trains the video diffusion model on class-specific datasets [7], [56], thereby lacking the generalization ability to handle generic scenes. Recently, CamCo [57] and CameraCtrl [58] introduced Plücker coordinates [59] in video diffusion models for camera motion control. Nevertheless, these methods still cannot precisely control the camera motion due to the complicated mapping from numeric camera parameters to videos. In this paper, we propose to leverage explicit point cloud representations for precise camera control in video generation, thereby fulfilling our needs for consistent and accurate novel view synthesis.

随着从多种输入模态生成内容的生成模型快速演进, 增强用户对生成过程的控制受到了广泛关注。ControlNet [44]、T2I-adapter [45] 和 GLIGEN [46] 率先为 T2I 生成引入了条件信号。类似策略也被用于视频生成, 可实现 RGB 图像 [17], [18], [47]、深度 [48], [49]、轨迹 [50], [51] 和语义图 [52] 等条件控制。然而, 相机运动控制受到的关注相对较少。AnimateDiff [53] 和 SVD [17] 研究类别条件的视频生成, 将相机运动分组, 并利用 LoRA [54] 模块生成特定的相机运动。MotionCtrl [13] 以相机外参矩阵作为条件信号, 改进了控制能力; 但它依赖一维数值, 虽对简单轨迹有效, 在复杂的真实场景中却难以精确控制。MultiDiff [55] 利用基于深度的图像变形生成变形图像, 并以变形图像为视频扩散模型的条件来提供显式三维先验; 然而它在特定类别的数据集 [7], [56] 上训练视频扩散模型, 因而缺乏处理通用场景的泛化能力。近来, CamCo [57] 和 CameraCtrl [58] 在视频扩散模型中引入普吕克坐标 [59] 来控制相机运动; 但由于从数值化的相机参数到视频之间的映射十分复杂, 这些方法仍无法精确控制相机运动。本文提出利用显式的点云表示来实现视频生成中的精确相机控制, 从而满足我们对一致且精确的新视角合成的需求。

3 方法 (Method)

In the following, we start with a brief introduction to video diffusion models in Section. 3.1, followed by an explanation of the point cloud reconstruction pipeline in Section. 3.2 and an illustration of the point-conditioned video diffusion model in Section. 3.3. Subsequently, we elaborate on the iterative view synthesis and camera trajectory planning strategy in Section. 3.4, and demonstrate how to apply our approach for

efficient 3D-GS optimization and text-to-3D generation in Section. 3.5.

下文中，我们先在第 3.1 节简要介绍视频扩散模型，随后在第 3.2 节说明点云重建流程，并在第 3.3 节阐述点云条件视频扩散模型。接着，我们在第 3.4 节详述迭代式视角合成与相机轨迹规划策略，并在第 3.5 节展示如何将本方法用于高效的 3D-GS 优化与文本到 3D 生成。

3.1 预备知识：视频扩散模型 (Preliminary: Video Diffusion Models)

A diffusion model [30] consists of two primary components: a forward process q and a reverse process p_θ . The forward process initiates with clean data $\mathbf{x}_0 \sim q_0(\mathbf{x}_0)$ and gradually introduces noise to $\mathbf{x}_0$, creating a noisy state across different time steps. This is mathematically represented as $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \epsilon$, where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The hyper-parameters α_t and σ_t satisfy the constraint $\alpha_t^2 + \sigma_t^2 = 1$. The reverse process p_θ focuses on removing noise from the clean data utilizing a noise predictor ϵ_θ , which is optimized by the objective:

扩散模型 [30] 主要由两个部分组成：前向过程 q 和反向过程 p_θ 。前向过程从干净数据 $\mathbf{x}_0 \sim q_0(\mathbf{x}_0)$ 出发，逐步向 $\mathbf{x}_0$ 注入噪声，在不同时间步形成含噪状态，数学上表示为 $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \epsilon$ ，其中 $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 。超参数 α_t 与 σ_t 满足约束 $\alpha_t^2 + \sigma_t^2 = 1$ 。反向过程 p_θ 则借助噪声预测器 ϵ_θ 从含噪数据中去除噪声，其优化目标为：

$$\min_{\theta} \mathbb{E}_{t \sim \mathcal{U}(0,1), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\|\epsilon_\theta(\mathbf{x}_t, t) - \epsilon\|_2^2]. \quad (1)$$

In diffusion-based video generation, Latent Diffusion Models (LDMs) [60] are frequently employed to mitigate the computational cost. In LDMs, the video data $\mathbf{x} \in \mathbb{R}^{L \times 3 \times H \times W}$ are encoded into the latent space using a pre-trained VAE encoder frame-by-frame, expressed as $\mathbf{z} = \mathcal{E}(\mathbf{x})$, $\mathbf{z} \in \mathbb{R}^{L \times C \times h \times w}$. Then, both the forward process q and the reverse process p_θ are performed in the latent space. The final generated videos are obtained through the VAE decoder $\hat{\mathbf{x}} = \mathcal{D}(\mathbf{z})$. In this work, we build our model based on an open-sourced Image-to-Video (I2V) diffusion model DynamiCrafter [18], which is capable of creating dynamic videos from a single input image. This aligns naturally with our goal of synthesizing novel views from sparse or single inputs.

在基于扩散的视频生成中，人们常采用潜在扩散模型 (Latent Diffusion Model, LDM) [60] 以降低计算开销。在 LDM 中，视频数据 $\mathbf{x} \in \mathbb{R}^{L \times 3 \times H \times W}$ 通过预训练的 VAE 编码器逐帧编码到潜在空间，记作 $\mathbf{z} = \mathcal{E}(\mathbf{x})$, $\mathbf{z} \in \mathbb{R}^{L \times C \times h \times w}$ 。随后，前向过程 q 与反向过程 p_θ 均在潜在空间中进行。最终生成的视频通过 VAE 解码器得到，即 $\hat{\mathbf{x}} = \mathcal{D}(\mathbf{z})$ 。本文中，我们的模型基于开源的图像到视频 (I2V) 扩散模型

DynamiCrafter [18] 构建，它能够从单张输入图像生成动态视频，这与我们从稀疏或单张输入合成新视角的目标天然契合。

3.2 从单张或稀疏图像重建点云 (Point Cloud Reconstruction from Single or Sparse Images)

To achieve accurate pose control in our novel view synthesis framework, we first establish the point cloud representation from the reference image(s). Specifically, we employ a dense stereo model, *e.g.* DUS3R [19], to reconstruct the point cloud and estimate camera parameters simultaneously. It takes a pair of RGB images $\mathbf{I}^0, \mathbf{I}^1 \in \mathbb{R}^{H \times W \times 3}$ as input and generates corresponding point maps $\mathbf{O}^{0,0}, \mathbf{O}^{1,0} \in \mathbb{R}^{H \times W \times 3}$ along with their respective confidence maps $\mathbf{D}^{0,0}, \mathbf{D}^{1,0} \in \mathbb{R}^{H \times W}$, reflecting the level of confidence in their accuracy. The subscript of $\mathbf{O}^{0,0}, \mathbf{O}^{1,0}$ denote that they are expressed in the same camera coordinate system of the anchor view $\mathbf{I}^0$. To recover the camera’s intrinsic parameters, it is assumed that the principal point is centered and the pixels are square. Consequently, only the focal length f_0^* remains unknown, which can be solved through a few optimization steps using Weiszfeld algorithm [61]:

为在新视角合成框架中实现精确的位姿控制，我们首先从参考图像构建点云表示。具体而言，我们采用稠密立体模型 (*e.g.* DUS3R [19]) 来同时重建点云并估计相机参数。它以一对 RGB 图像 $\mathbf{I}^0, \mathbf{I}^1 \in \mathbb{R}^{H \times W \times 3}$ 为输入，生成对应的点图 $\mathbf{O}^{0,0}, \mathbf{O}^{1,0} \in \mathbb{R}^{H \times W \times 3}$ ，以及各自的置信度图 $\mathbf{D}^{0,0}, \mathbf{D}^{1,0} \in \mathbb{R}^{H \times W}$ ，后者反映点图精度的置信程度。 $\mathbf{O}^{0,0}, \mathbf{O}^{1,0}$ 的下标表示它们都表达在锚定视角 $\mathbf{I}^0$ 的同一相机坐标系中。为恢复相机内参，这里假设主点位于中心且像素为正方形，于是只剩焦距 f_0^* 未知，可用 Weiszfeld 算法 [61] 经过几步优化求解：

$$f_0^* = \arg \min_{f_0} \sum_{u=0}^W \sum_{v=0}^H \mathbf{D}_{u,v}^{0,0} \left\| (u', v') - f_0 \frac{(\mathbf{O}_{u,v,0}^{0,0}, \mathbf{O}_{u,v,1}^{0,0})}{\mathbf{O}_{u,v,2}^{0,0}} \right\|, \quad (2)$$

where $u' = u - \frac{W}{2}$ and $v' = v - \frac{H}{2}$. In the case where only a single input image is available, we duplicate the input image to create a paired input and then estimate its point map and camera intrinsic. When there are more than two input images, it can also perform global point map alignment with a few optimization iterations.

其中 $u' = u - \frac{W}{2}$, $v' = v - \frac{H}{2}$ 。当仅有单张输入图像时，我们把该图像复制一份构成成对输入，再估计其点图与相机内参。当输入图像多于两张时，它还能通过几步优化迭代实现全局点图对齐。

The colored point cloud, which provides coarse 3D information of the scene, can be obtained by integrating the point maps with their corresponding RGB images. However, the limited representation capabilities of the point cloud and

the insufficient 3D cues provided by sparse or single inputs can result in significant missing regions, occlusions, and artifacts in the reconstructed point cloud, leading to low-quality render results. Therefore, we propose incorporating video diffusion models to achieve high-fidelity novel view synthesis based on the imperfect point cloud.

将点图与对应的 RGB 图像结合, 即可得到彩色点云, 为场景提供粗略的三维信息。然而, 点云表示能力有限, 加之稀疏或单张输入所提供的三维线索不足, 重建出的点云会出现明显的缺失区域、遮挡和伪影, 导致渲染结果质量低下。因此, 我们提出引入视频扩散模型, 在这种不完美的点云基础上实现高保真新视角合成。

3.3 用视频扩散模型渲染高保真新视角 (Rendering High-fidelity Novel Views with Video Diffusion Models)

As shown in Fig. 1, taking a single reference image $\mathbf{I}^{\text{ref}}$ as an example, we first obtain its point cloud, camera intrinsics and camera pose $\mathbf{C}^{\text{ref}}$ through the dense stereo model [19]. Subsequently, we can navigate the camera along a camera pose sequence $\mathbf{C} = \{\mathbf{C}^0, \dots, \mathbf{C}^{L-1}\}$ that contains $\mathbf{C}^{\text{ref}}$ to render the point cloud and obtain a sequence of render results, denote as $\mathbf{P} = \{\mathbf{P}^0, \dots, \mathbf{P}^{L-1}\}$. While the point cloud renders accurately represent view relationships, they are plagued by substantial occlusions, missing areas, and reduced visual fidelity. To enhance the quality of novel view rendering, our objective is to learn a conditional distribution $\mathbf{x} \sim p(\mathbf{x} | \mathbf{I}^{\text{ref}}, \mathbf{P})$ that can produce high-quality novel views $\mathbf{x} = \{\mathbf{x}^0, \dots, \mathbf{x}^{L-1}\}$ based on the point cloud renders $\mathbf{P}$ and the reference image(s) $\mathbf{I}^{\text{ref}}$. Motivated by the efficacy of video diffusion models [16], [17], [18] in synthesizing high-quality and consistent videos, we learn this conditional distribution by training a video diffusion model conditioned on the point cloud renders and the reference image(s). As a result, the novel view synthesis process can be naturally modeled as the reverse process of a point-conditioned video diffusion model, expressed as $\mathbf{x} \sim p_{\theta}(\mathbf{x} | \mathbf{I}^{\text{ref}}, \mathbf{P})$, where θ denotes the model parameters.

如图 1 所示, 以单张参考图像 $\mathbf{I}^{\text{ref}}$ 为例, 我们先通过稠密立体模型 [19] 得到其点云、相机内参和相机位姿 $\mathbf{C}^{\text{ref}}$ 。随后, 我们可以让相机沿一条包含 $\mathbf{C}^{\text{ref}}$ 的相机位姿序列 $\mathbf{C} = \{\mathbf{C}^0, \dots, \mathbf{C}^{L-1}\}$ 移动, 渲染点云并得到一系列渲染结果, 记作 $\mathbf{P} = \{\mathbf{P}^0, \dots, \mathbf{P}^{L-1}\}$ 。点云渲染虽能准确体现视角间的关系, 却饱受大量遮挡、缺失区域和视觉保真度下降之扰。为提升新视角渲染质量, 我们的目标是学习一个条件分布 $\mathbf{x} \sim p(\mathbf{x} | \mathbf{I}^{\text{ref}}, \mathbf{P})$, 使其能够根据点云渲染 $\mathbf{P}$ 和参考图像 $\mathbf{I}^{\text{ref}}$ 生成高质量的新视角 $\mathbf{x} = \{\mathbf{x}^0, \dots, \mathbf{x}^{L-1}\}$ 。鉴于视频扩散模型 [16], [17], [18] 在合成高质量、一致视频上的成效, 我们通过训练一个以点云渲染和参考图像为条件的视频扩散模型来

学习该条件分布。于是, 新视角合成过程可自然地建模为点云条件视频扩散模型的反向过程, 表示为 $\mathbf{x} \sim p_{\theta}(\mathbf{x} | \mathbf{I}^{\text{ref}}, \mathbf{P})$, 其中 θ 表示模型参数。

The architecture of the point-conditioned video diffusion model is illustrated in Fig. 1. It inherits the LDM [60] architecture, which primarily comprises a pair of VAE encoder $\mathcal{E}$ and decoder $\mathcal{D}$ for image compression, a video denoising U-Net with spatial layers followed by temporal layers for temporal-aware noise estimation, as well as a CLIP [62] image encoder for reference image understanding. We incorporate point cloud renders as conditional signals in the video denoising U-Net by encoding them using $\mathcal{E}$ and concatenating the resulting latent images with noise across the channel dimension.

点云条件视频扩散模型的架构如图 1 所示。它沿用 LDM [60] 架构, 主要包括: 用于图像压缩的一对 VAE 编码器 $\mathcal{E}$ 与解码器 $\mathcal{D}$; 一个视频去噪 U-Net, 其空间层之后接时序层, 用于时序感知的噪声估计; 以及一个用于理解参考图像的 CLIP [62] 图像编码器。我们把点云渲染作为条件信号加入视频去噪 U-Net: 先用 $\mathcal{E}$ 对其编码, 再将得到的潜在图像与噪声沿通道维拼接。

To train the model, we create paired training data that includes both point cloud renders $\mathbf{P} = \{\mathbf{P}^0, \dots, \mathbf{P}^{L-1}\}$ and the corresponding ground-truth reference images $\mathbf{I} = \{\mathbf{I}^0, \dots, \mathbf{I}^{L-1}\}$. The point cloud renders are forced to traverse at least one ground-truth view, *i.e.*, to include at least one ground-truth reference image at a random location among the L frames. It helps the model better learn to transfer fine details from the reference image(s) to the point cloud renders and enables our model to flexibly handle arbitrary number of reference image(s). Following the approach of LDMs [60], we freeze the parameters of the VAE encoder $\mathcal{E}$ and decoder $\mathcal{D}$, and conduct the training process in the latent space. Specifically, we encode the training data pair $\mathbf{I} = \{\mathbf{I}^0, \dots, \mathbf{I}^{L-1}\}$ and $\mathbf{P} = \{\mathbf{P}^0, \dots, \mathbf{P}^{L-1}\}$ into the latent space, yielding the ground-truth latents $\mathbf{z} = \{\mathbf{z}^0, \dots, \mathbf{z}^{L-1}\}$ and the condition signals $\hat{\mathbf{z}} = \{\hat{\mathbf{z}}^0, \dots, \hat{\mathbf{z}}^{L-1}\}$ that will be concatenated channel-wise with the sampled noise. Subsequently, the video denoising U-Net is optimized by the diffusion loss:

为训练模型, 我们构造成对训练数据, 既包含点云渲染 $\mathbf{P} = \{\mathbf{P}^0, \dots, \mathbf{P}^{L-1}\}$, 也包含对应的真值参考图像 $\mathbf{I} = \{\mathbf{I}^0, \dots, \mathbf{I}^{L-1}\}$ 。我们强制点云渲染至少经过一个真值视角, 即在 L 帧中的某个随机位置至少包含一张真值参考图像。这有助于模型更好地学会把参考图像中的细节迁移到点云渲染上, 并使模型能够灵活处理任意数量的参考图像。遵循 LDM [60] 的做法, 我们冻结 VAE 编码器 $\mathcal{E}$ 与解码器 $\mathcal{D}$ 的参数, 在潜在空间中开展训练。具体而言, 我们将训练数据对 $\mathbf{I} = \{\mathbf{I}^0, \dots, \mathbf{I}^{L-1}\}$ 与 $\mathbf{P} = \{\mathbf{P}^0, \dots, \mathbf{P}^{L-1}\}$ 编码到潜在空间, 得到真值潜在表示 $\mathbf{z} = \{\mathbf{z}^0, \dots, \mathbf{z}^{L-1}\}$ 以及将与采样噪声沿

通道拼接的条件信号 $\hat{z} = \{\hat{z}^0, \dots, \hat{z}^{L-1}\}$ 。随后，视频去噪 U-Net 由扩散损失优化：

$$\min_{\theta} = \mathbb{E}_{t \sim \mathcal{U}(0,1), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\|\epsilon_{\theta}(\mathbf{z}_t, t, \hat{z}, \mathbf{I}^{\text{ref}}) - \epsilon\|_2^2], \quad (3)$$

where $\mathbf{z}_t = \alpha_t \mathbf{z}_0 + \sigma_t \epsilon$.

其中 $\mathbf{z}_t = \alpha_t \mathbf{z}_0 + \sigma_t \epsilon$ 。

During the inference process, we render a sequence of point cloud renders $\mathbf{P} = \{\mathbf{P}^0, \dots, \mathbf{P}^{L-1}\}$ and replace the reference view render results with the corresponding reference image(s). Subsequently, we encode them into the latent space to obtain the latent images $\hat{z} = \{\hat{z}^0, \dots, \hat{z}^{L-1}\}$, sample noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then concatenate them channel-wise to construct the noisy latent. In addition to the latent space condition, we also pass the reference image(s) into the CLIP image [62] encoder, which will modulate the U-Net features through cross-attention for better 3D understanding. With the trained U-Net, the noisy latents are iteratively denoised into clean latents and then decoded into high-quality novel views $\mathbf{x} = \{\mathbf{x}^0, \dots, \mathbf{x}^{L-1}\}$ using the VAE decoder $\mathcal{D}$.

在推理过程中，我们渲染出一列点云渲染 $\mathbf{P} = \{\mathbf{P}^0, \dots, \mathbf{P}^{L-1}\}$ ，并把参考视角的渲染结果替换为对应的参考图像。随后将它们编码到潜在空间，得到潜在图像 $\hat{z} = \{\hat{z}^0, \dots, \hat{z}^{L-1}\}$ ，采样噪声 $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ，再沿通道拼接构成含噪潜在表示。除潜在空间条件外，我们还把参考图像送入 CLIP [62] 图像编码器，通过交叉注意力调制 U-Net 特征，以获得更好的三维理解。借助训练好的 U-Net，含噪潜在表示被迭代去噪为干净潜在表示，再由 VAE 解码器 $\mathcal{D}$ 解码为高质量的新视角 $\mathbf{x} = \{\mathbf{x}^0, \dots, \mathbf{x}^{L-1}\}$ 。

3.4 迭代式视角合成与相机轨迹规划 (Iterative View Synthesis and Camera Trajectory Planning)

Existing video diffusion models encounter challenges in generating long videos with numerous frames. As the video length increases during inference, it results in decreased video stability and increased computational costs. Therefore, the trained ViewCrafter model may face challenges in generating longer videos to produce a larger view range. To address this challenge, we adopt an iterative view synthesis approach. Specifically, given an initial point cloud established from the reference image(s), we navigate the camera from one of the reference views to a target camera pose to reveal occlusion and missing regions of the current point cloud. Subsequently, we can generate high-fidelity novel views using ViewCrafter and back-project the generated views to complete the point cloud. Through iteratively moving the camera, generating novel views, and updating the point cloud, we can ultimately obtain novel views with an extended view range and a more complete point cloud representation of the scene.

算法 1 相机轨迹规划

输入： 参考图像 $\mathcal{I}_{\text{ref}}$ ，稠密立体模型 $\mathcal{D}(\cdot)$ ，点云条件视频扩散模型 $\mathcal{V}(\cdot)$ ，初始点云 $\mathcal{P}_{\text{ref}}$ ，搜索空间 $\mathcal{S}$ ，初始位姿 $\mathcal{C}_{\text{ref}}$ ，最大预测位姿数 N ，候选位姿数 K ，效用函数 $\mathcal{F}(\cdot)$

- 1: **初始化** 当前点云 $\mathcal{P}_{\text{curr}} \leftarrow \mathcal{P}_{\text{ref}}$, 当前相机位姿 $\mathcal{C}_{\text{curr}} \leftarrow \mathcal{C}_{\text{ref}}$, $step \leftarrow 0$
- 2: **while** $step \leq N$ **do**
- 3: 从当前位姿 $\mathcal{C}_{\text{curr}}$ 周围的搜索空间 $\mathcal{S}$ 中均匀采样 K 个候选位姿 $\mathcal{C}_{\text{can}} = \{\mathcal{C}_{\text{can}}^1, \dots, \mathcal{C}_{\text{can}}^K\}$, 初始化候选掩码集合 $\mathcal{M}_{\text{can}} = \{\}$
- 4: **for** $\mathcal{C}$ in $\{\mathcal{C}_{\text{can}}^1, \dots, \mathcal{C}_{\text{can}}^K\}$ **do**
- 5: $\mathcal{M}_{\mathcal{C}} = \text{Render}(\mathcal{P}_{\text{curr}}, \mathcal{C})$
- 6: $\mathcal{M}_{\text{can}}.\text{append}(\mathcal{M}_{\mathcal{C}})$
- 7: **end for**
- 8: $\mathcal{C}_{\text{nbv}} = \arg \max_{\mathcal{C} \in \mathcal{C}_{\text{can}}} \mathcal{F}(\mathcal{C})$
- 9: $\mathcal{I}_{\text{nbv}} = \mathcal{V}(\text{interpolate}(\mathcal{C}_{\text{curr}}, \mathcal{C}_{\text{nbv}}), \mathcal{P}_{\text{curr}})$
- 10: $\mathcal{P}_{\text{curr}} \leftarrow \mathcal{D}(\mathcal{I}_{\text{nbv}}, \mathcal{P}_{\text{curr}})$
- 11: $\mathcal{C}_{\text{curr}} \leftarrow \mathcal{C}_{\text{nbv}}$
- 12: $step \leftarrow step + 1$
- 13: **end while**
- 14: **返回**

现有视频扩散模型在生成帧数众多的长视频时面临困难。推理时随着视频长度增加，视频稳定性下降、计算开销上升。因此，训练好的 ViewCrafter 模型在生成更长视频以覆盖更大视角范围时也会遇到困难。为应对这一挑战，我们采用迭代式视角合成方法。具体来说，给定由参考图像建立的初始点云，我们把相机从某个参考视角移动到一个目标相机位姿，以揭示当前点云的遮挡和缺失区域；随后用 ViewCrafter 生成高保真新视角，并将生成的视角反投影回去以补全点云。通过反复移动相机、生成新视角、更新点云，我们最终能得到视角范围更大的新视角，以及更完整的场景点云表示。

In the iterative view synthesis process, the design of the camera trajectory significantly impacts the synthesis results. Methods like [14], [63] use predefined camera trajectories for scene generation, which overlooks the diverse geometry relationships presented in different scenes, resulting in significant occlusions. To effectively reveal occlusions in the iterative view synthesis process and facilitate more complete scene generation, we designed a Next-Best-View (NBV) [64], [65], [66] - based camera trajectory planning algorithm, which enables the adaptive generation of camera trajectories tailored to handle various scene types. The camera trajectory planning algorithm is illustrated in Algorithm. 1. Starting with the input reference image(s) $\mathcal{I}_{\text{ref}}$, we construct an initial point cloud $\mathcal{P}_{\text{ref}}$ using the dense stereo model [19]. Referring [64], [66], [67], [68], we

opt for a forward-facing quarter-sphere with evenly distributed camera poses as the searching space, denoted as $\mathcal{S}$, and position it centrally at the origin of the point cloud’s world coordinate system, setting the radius to the depth of the center pixel in the reference image. The camera trajectory is initialized from one of the reference camera poses $\mathcal{C}_{\text{ref}}$. To predict the subsequent pose, we uniformly sample K candidate camera poses $\mathcal{C}_{\text{can}} = \{\mathcal{C}_{\text{can}}^1, \dots, \mathcal{C}_{\text{can}}^K\}$ from the searching space surrounding the current camera pose $\mathcal{C}_{\text{curr}} = \mathcal{C}_{\text{ref}}$, then render a set of candidate masks $\mathcal{M}_{\text{can}}$ (where 1 signifies occlusion and missing regions, while 0 represents filled regions) from the current point cloud $\mathcal{P}_{\text{curr}}$. We then establish a utility function [64] $\mathcal{F}(\cdot)$ to determine the optimal camera pose for the subsequent step, defined as:

在迭代式视角合成过程中，相机轨迹的设计对合成结果影响显著。[14], [63] 等方法采用预定义的相机轨迹进行场景生成，忽视了不同场景中多样的几何关系，导致明显的遮挡。为在迭代式视角合成中有效揭示遮挡、促成更完整的场景生成，我们设计了一种基于下一最佳视角（Next-Best-View, NBV）[64], [65], [66] 的相机轨迹规划算法，能够自适应地生成契合各类场景的相机轨迹。该相机轨迹规划算法如算法 1 所示。从输入参考图像 $\mathcal{I}_{\text{ref}}$ 出发，我们用稠密立体模型 [19] 构建初始点云 $\mathcal{P}_{\text{ref}}$ 。参照 [64], [66], [67], [68]，我们选取一个相机位姿均匀分布、朝向前方的四分之一球面作为搜索空间，记作 $\mathcal{S}$ ，并把它的中心放在点云世界坐标系的原点，半径设为参考图像中心像素的深度。相机轨迹从某个参考相机位姿 $\mathcal{C}_{\text{ref}}$ 初始化。为预测下一个位姿，我们在当前相机位姿 $\mathcal{C}_{\text{curr}} = \mathcal{C}_{\text{ref}}$ 周围的搜索空间中均匀采样 K 个候选相机位姿 $\mathcal{C}_{\text{can}} = \{\mathcal{C}_{\text{can}}^1, \dots, \mathcal{C}_{\text{can}}^K\}$ ，然后从当前点云 $\mathcal{P}_{\text{curr}}$ 渲染出一组候选掩码 $\mathcal{M}_{\text{can}}$ （其中 1 表示遮挡与缺失区域，0 表示已填充区域）。接着，我们建立一个效用函数 [64] $\mathcal{F}(\cdot)$ 来确定下一步的最优相机位姿，定义为：

$$\mathcal{F}(\mathcal{C}) = \begin{cases} \frac{\text{sum}(\mathcal{M}_{\mathcal{C}})}{W \times H}, \frac{\text{sum}(\mathcal{M}_{\mathcal{C}})}{W \times H} < \Theta \\ 1 - \frac{\text{sum}(\mathcal{M}_{\mathcal{C}})}{W \times H}, \frac{\text{sum}(\mathcal{M}_{\mathcal{C}})}{W \times H} > \Theta, \end{cases} \quad (4)$$

where $\mathcal{C} \in \mathcal{C}_{\text{can}}$, $\mathcal{M}_{\mathcal{C}} \in \mathcal{M}_{\text{can}}$, and $\text{sum}(\mathcal{M}_{\mathcal{C}}) = \sum_{u=0}^W \sum_{v=0}^H \mathcal{M}_{\mathcal{C}}(u, v)$. The utility function helps identify a suitable camera pose that reveals an adequate area of occlusion and missing regions while avoiding poses that reveal excessively large holes deviating significantly from a threshold Θ , which may affect ViewCrafter’s generation capability. Once the next best camera pose $\mathcal{C}_{\text{nbv}}$ is predicted, we interpolate a camera path between $\mathcal{C}_{\text{curr}}$ and $\mathcal{C}_{\text{nbv}}$, and then apply ViewCrafter along the path to generate a sequence of high-fidelity novel views. Subsequently, we back-project and align the generated novel view $\mathcal{I}_{\text{nbv}}$ onto the current point cloud $\mathcal{P}_{\text{curr}}$, and designate $\mathcal{C}_{\text{nbv}}$ as the new $\mathcal{C}_{\text{curr}}$, then repeat the aforementioned process until the predicted poses reach the predefined limitation N . Through iteratively predicting camera poses, synthesizing novel views,

and updating the point cloud, we can ultimately obtain a more complete point cloud representation of the scene.

其中 $\mathcal{C} \in \mathcal{C}_{\text{can}}$, $\mathcal{M}_{\mathcal{C}} \in \mathcal{M}_{\text{can}}$, 且 $\text{sum}(\mathcal{M}_{\mathcal{C}}) = \sum_{u=0}^W \sum_{v=0}^H \mathcal{M}_{\mathcal{C}}(u, v)$ 。该效用函数有助于挑选一个合适的相机位姿：既能揭示足够面积的遮挡与缺失区域，又能避免那些揭示出的空洞过大、明显偏离阈值 Θ 的位姿（这类位姿可能影响 ViewCrafter 的生成能力）。一旦预测出下一最佳相机位姿 $\mathcal{C}_{\text{nbv}}$ ，我们便在 $\mathcal{C}_{\text{curr}}$ 与 $\mathcal{C}_{\text{nbv}}$ 之间插值出一条相机路径，并沿该路径运行 ViewCrafter 生成一系列高保真新视角。随后，我们将生成的新视角 $\mathcal{I}_{\text{nbv}}$ 反投影并对齐到当前点云 $\mathcal{P}_{\text{curr}}$ 上，把 $\mathcal{C}_{\text{nbv}}$ 指定为新的 $\mathcal{C}_{\text{curr}}$ ，再重复上述过程，直到预测的位姿数达到预设上限 N 。通过反复预测相机位姿、合成新视角、更新点云，我们最终能得到更完整的场景点云表示。

3.5 应用 (Applications)

ViewCrafter can effectively produce accurate, consistent, and high-fidelity novel views from single or sparse inputs. Nevertheless, it faces challenges in providing immersive experiences due to the slow multi-step denoising process. To achieve real-time rendering, we further delve into optimizing a 3D-GS [2] representation from the results of our ViewCrafter. To that aim, a direct approach involves concurrently running ViewCrafter multiple times on the initially built point cloud to generate multiple novel views and optimizing a 3D-GS from them. However, this will lead to suboptimal optimization results, since the initial point cloud is incomplete and will introduce inconsistencies in occlusion regions among the generated view sequences at different times.

ViewCrafter 能够从单张或稀疏输入有效生成精确、一致、高保真的新视角。然而，由于多步去噪过程较慢，它在提供沉浸式体验上仍有困难。为实现实时渲染，我们进一步深入研究如何从 ViewCrafter 的结果优化出一个 3D-GS [2] 表示。为此，一种直接的做法是在初始构建的点云上并发多次运行 ViewCrafter 生成多组新视角，再据此优化 3D-GS。但由于初始点云并不完整，不同次生成的视角序列在遮挡区域会引入不一致，这种做法会导致次优的优化结果。

As illustrated in Fig. 2, to facilitate more consistent 3D-GS optimization, we leverage the aforementioned iterative view synthesis strategy to iteratively complete the initial point cloud and synthesize novel views using ViewCrafter, which not only provides consistent novel views as training data but also offers a strong geometry initialization for the 3D-GS [2]. During training, the center of each 3D Gaussian is initialized from the completed dense point cloud, and the attributes of each 3D Gaussian are optimized under the supervision of the synthesized novel views. We simplify the 3D-GS optimization process by deprecating the densification, splitting, and opacity reset

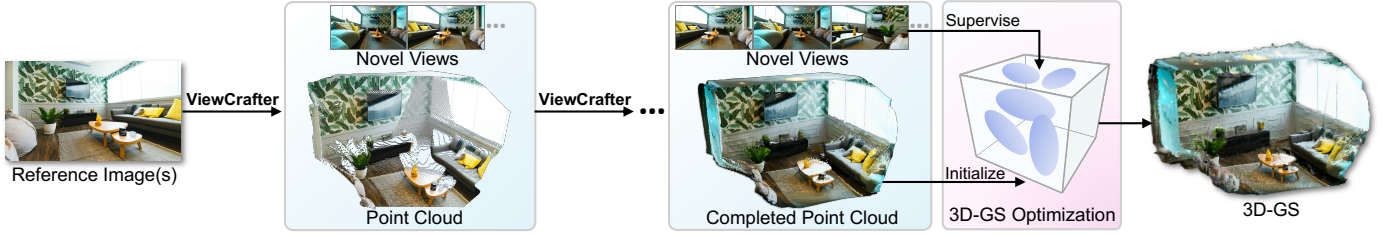


图 2. 3D-GS 优化的应用。为促成更一致的 3D-GS 优化，我们利用迭代式视角合成策略，用 ViewCrafter 逐步补全初始点云并合成新视角。随后，我们用补全后的稠密点云初始化 3D-GS，并用合成的新视角监督 3D-GS 的训练。**Application of 3D-GS optimization.** To facilitate more consistent 3D-GS optimization, we leverage the iterative view synthesis strategy to progressively complete the initial point cloud and synthesize novel views using ViewCrafter. We then use the completed dense point cloud to initialize 3D-GS and employ the synthesized novel views to supervise 3D-GS training.

tricks [69], and reduce the overall optimization time into 2,000 iterations, which is considerably faster than the original 3D-GS training.

如图 2 所示，为促成更一致的 3D-GS 优化，我们利用前述迭代式视角合成策略，用 ViewCrafter 迭代地补全初始点云并合成新视角。这样既提供了一致的新视角作为训练数据，也为 3D-GS [2] 提供了强有力的几何初始化。训练时，每个三维高斯的中心由补全后的稠密点云初始化，其各项属性则在合成新视角的监督下优化。我们简化了 3D-GS 优化流程，弃用了稠密化、分裂和不透明度重置等技巧 [69]，并将总优化时长压缩到 2000 次迭代，比原始 3D-GS 训练快得多。

In addition to synthesizing novel views of real-world images, we also explore the application of combining ViewCrafter with creative text-to-image diffusion models for text-to-3D generation. This involves using a text-to-image diffusion model to generate a reference image from the provided text prompt, followed by employing ViewCrafter for novel view synthesis and 3D reconstruction.

除对真实世界图像合成新视角外，我们还探索将 ViewCrafter 与富有创意的文本到图像扩散模型结合，用于文本到 3D 生成。其做法是：先用文本到图像扩散模型根据给定的文本提示生成一张参考图像，再用 ViewCrafter 进行新视角合成与三维重建。

4 实验 (Experiments)

In this section, we begin with an illustration of the implementation details in Section. 4.1, followed by a comparison of zero-shot novel view synthesis in Section. 4.2 and scene reconstruction in Section. 4.3. Subsequently, we conduct ablation studies to evaluate design choices and training strategies in Section. 4.4. Finally, we present the results of text-to-3D generation in Section. 4.5.

本节中，我们先在第 4.1 节说明实现细节，随后在第 4.2 节比较零样本新视角合成、在第 4.3 节比较场景重建。接着，我们在第 4.4 节开展消融研究，评估各项设计选择与训练策略。最后，我们在第 4.5 节展示文本到 3D 生成的结果。

4.1 实现细节 (Implementation Details)

We employ a progressive training strategy. In the first stage, we train the ViewCrafter model at a resolution of 320×512 , with the frame length set to 25. The entire video denoising U-Net is trained for 50,000 iterations using a learning rate of 5×10^{-5} and a mini-batch size of 16. In the second stage, we fine-tune the spatial layers (*i.e.*, 2D Conv and spatial attention layers) of the video denoising U-Net at a resolution of 576×1024 for high-resolution adaptation, with 5,000 iterations on a learning rate of 1×10^{-5} and a valid mini-batch size of 16. Our model was trained on a mixed dataset consisting of DL3DV [70] and RealEstate10K [7]. We divide the video data into video clips, each containing 25 frames. To generate the conditional signals, specifically the point cloud renders, we process the video clips using DUST3R [19] to obtain the camera trajectory of the video clips and the globally aligned point clouds of each video frame. Then, we randomly select the constructed point cloud of the video frames and render it along the estimated camera trajectory using Pytorch3D [71]. In total, we generate 632,152 video pairs as training data. During inference, we adopt DDIM sampler [30] with classifier-free guidance [72].

我们采用渐进式训练策略。第一阶段，我们在 320×512 分辨率下训练 ViewCrafter 模型，帧长设为 25。整个视频去噪 U-Net 训练 50,000 次迭代，学习率为 5×10^{-5} ，小批量大小为 16。第二阶段，为适配高分辨率，我们在 576×1024 分辨率下微调视频去噪 U-Net 的空间层（即 2D 卷积层与空间注意力层），迭代 5,000 次，学习率为 1×10^{-5} ，有效小批量大小为 16。我们的模型在由 DL3DV [70] 和 RealEstate10K [7] 组成的混合数据集上训练。我们将视频数据切分为视频片



图 3. Tanks-and-Temples [21], RealEstate10K [7] 和 CO3D [39] 数据集上零样本新视角合成的定性对比。参考图像位于最左列，真值新视角位于最右列。Qualitative comparison of zero-shot novel view synthesis on Tanks-and-Temples [21], RealEstate10K [7] and CO3D [39] dataset. The reference images are displayed in the left-most column, and the ground truth novel views are located in the right-most column.

段，每段含 25 帧。为生成条件信号（即点云渲染），我们用 DUST3R [19] 处理视频片段，得到片段的相机轨迹以及每一视频帧全局对齐后的点云。然后，我们随机选取若干视频帧所构建的点云，用 Pytorch3D [71] 沿估计出的相机轨迹渲染。总计生成 632,152 对视频对作为训练数据。推理时，我们采用带无分类器引导 [72] 的 DDIM 采样器 [30]。

4.2 零样本新视角合成对比 (Zero-shot Novel View Synthesis Comparison)

数据集与评价指标 (Datasets and evaluation metrics). In our study, we employ three real-world datasets of different scales as our zero-shot novel view synthesis evaluation benchmark, which includes the CO3D [39] dataset, the RealEstate10K [7] dataset, and the Tanks-and-Temples [21] dataset. For CO3D [39] consisting of object-centric scenes, we evaluate on 10 scenes. RealEstate10K [7] comprises video clips of indoor scenes, we adopt 10 scenes from its test set for evalua-

表 1

Tanks-and-Temples [21]、RealEstate10K [7] 和 CO3D [39] 数据集上零样本新视角合成的定量对比。由于 ZeroNVS [12] 和 LucidDreamer [14] 只能处理正方形图像，我们在计算定量指标时，把本方法和 MotionCtrl [13] 生成的新视角裁剪至与它们对齐。**Quantitative comparison of zero-shot novel view synthesis on Tanks-and-Temples [21], RealEstate10K [7] and CO3D [39] dataset.** Since ZeroNVS [12] and LucidDreamer [14] can only handle square images, we crop the generated novel views from our method and MotionCtrl [13] to align with them when computing the quantitative metrics.

数据集	简单集						困难集					
方法	LPIPS ↓	PSNR ↑	SSIM ↑	FID ↓	R_{dist} ↓	T_{dist} ↓	LPIPS ↓	PSNR ↑	SSIM ↑	FID ↓	R_{dist} ↓	T_{dist} ↓
Tanks-and-Temples												
LucidDreamer [14]	0.413	14.53	0.362	42.32	6.137	5.695	0.558	11.69	0.267	200.8	8.998	9.305
ZeroNVS [12]	0.482	14.71	0.380	74.60	8.810	6.348	0.569	12.05	0.309	131.0	8.860	8.557
MotionCtrl [13]	0.400	15.34	0.427	70.3	7.299	8.039	0.473	13.29	0.384	196.8	9.801	9.112
ViewCrafter (本文)	0.194	21.26	0.655	27.18	0.471	1.009	0.283	18.07	0.563	38.92	1.109	0.910
RealEstate10K												
LucidDreamer [14]	0.315	16.35	0.579	56.77	5.821	10.02	0.400	14.13	0.511	71.43	7.990	10.85
ZeroNVS [12]	0.364	16.50	0.577	96.18	6.370	9.817	0.431	14.24	0.535	105.8	8.562	10.31
MotionCtrl [13]	0.341	16.31	0.604	89.90	4.236	9.091	0.386	16.29	0.587	70.02	8.084	9.295
ViewCrafter (本文)	0.145	21.81	0.796	33.09	0.380	2.888	0.178	22.04	0.798	24.89	1.098	2.867
CO3D												
LucidDreamer [14]	0.429	15.11	0.451	78.87	12.90	6.665	0.517	12.69	0.374	157.8	16.43	8.301
ZeroNVS [12]	0.467	15.15	0.463	93.84	15.44	8.872	0.524	13.31	0.426	143.2	15.02	10.22
MotionCtrl [13]	0.393	16.87	0.529	69.18	16.87	5.131	0.443	15.46	0.502	112.7	18.81	5.575
ViewCrafter (本文)	0.243	21.38	0.687	24.63	2.175	1.033	0.324	18.96	0.641	36.96	2.849	1.480

tion. For Tanks-and-Temples [21] containing large-scale outdoor and indoor scenes, we use all of the 9 scenes. For all benchmarks, we extract frames from the original captured videos and create two types of test sets by applying different sampling rates to the original video. The easy test set is generated using a small frame sampling stride, characterized by slow camera motions and limited view ranges. In contrast, the hard test set is produced with a large sampling stride, featuring rapid camera motions and large view ranges.

在研究中，我们采用三个不同规模的真实世界数据集作为零样本新视角合成的评测基准，包括 CO3D [39]、RealEstate10K [7] 和 Tanks-and-Temples [21] 数据集。对于以物体为中心的 CO3D [39]，我们在 10 个场景上评测。RealEstate10K [7] 由室内场景视频片段构成，我们从其测试集中选取 10 个场景评测。对于包含大规模室外与室内场景的 Tanks-and-Temples [21]，我们使用全部 9 个场景。对所有基准，我们都从原始拍摄视频中抽取帧，并对原视频施加不同的采样率，构造两类测试集。简单 (easy) 测试集用较小的抽帧步长生成，相机运动缓慢、视角范围有限；困难 (hard) 测试集则用较大的采样步长生成，相机运动快速、视角范围大。

We employ PSNR, SSIM [73], LPIPS [74], and FID [75] as the evaluation metrics for assessing image quality. Among

these, PSNR is a traditional metric used to compare image similarity. SSIM [73] and LPIPS [74] are designed to evaluate the structural and perceptual similarity between the generated images and the ground truth images, as these metrics are specifically designed to align more closely with human perceptual judgment. Referring to [3], [4], [5], we further integrate FID into our evaluation process for assessing the quality of synthesized views, which proves particularly efficacious when evaluating the hard test set that contains a significant number of missing and occlusion regions. Additionally, to evaluate the pose accuracy of the generated novel views, we estimate the camera poses of the generated novel views to compare with the ground truth camera poses. Following [58], we transform the camera coordinate of the estimated poses to be relative to the first frame, and normalize the translation scale using the furthest frame. We then calculate the rotation distance (R_{dist}) in comparison to the ground truth rotation matrices of each generated novel view sequence, expressed as:

我们采用 PSNR、SSIM [73]、LPIPS [74] 和 FID [75] 作为评估图像质量的指标。其中，PSNR 是用于比较图像相似度的传统指标；SSIM [73] 和 LPIPS [74] 则用于评估生成图像与真值图像之间的结构相似度和感知相似度，这两个指标专门设计得与人类感知判断更为吻合。参照 [3], [4], [5]，我们进

一步把 FID 纳入评估流程来衡量合成视角的质量；在评估含有大量缺失与遮挡区域的困难测试集时，FID 尤为有效。此外，为评估生成新视角的位姿精度，我们估计生成新视角的相机位姿并与真值相机位姿比较。遵循 [58]，我们把估计位姿的相机坐标变换为相对于第一帧，并用最远帧对平移尺度归一化。随后，我们对每条生成的新视角序列，计算其与真值旋转矩阵之间的旋转距离 (R_{dist})，表示为：

$$R_{\text{dist}} = \sum_{i=1}^n \arccos\left(\frac{\text{tr}(\mathbf{R}_{\text{gen}}^i \mathbf{R}_{\text{gt}}^{iT}) - 1}{2}\right), \quad (5)$$

where $\mathbf{R}_{\text{gt}}^i$ and $\mathbf{R}_{\text{gen}}^i$ denote the ground truth rotation matrix and generated rotation matrix, and we sum the distance of all frames as the final results. We also compute the translation distance (T_{dist}), expressed as:

其中 $\mathbf{R}_{\text{gt}}^i$ 与 $\mathbf{R}_{\text{gen}}^i$ 分别表示真值旋转矩阵和生成旋转矩阵，我们将所有帧的距离求和作为最终结果。我们还计算平移距离 (T_{dist})，表示为：

$$T_{\text{dist}} = \sum_{i=1}^n \|\mathbf{T}_{\text{gt}}^i - \mathbf{T}_{\text{gen}}^i\|_2. \quad (6)$$

Notably, since COLMAP [76] is sensitive to inconsistent features and prone to fail to extract poses from the generated novel views, we instead use DUST3R [19] for more robust pose estimation.

值得注意的是，由于 COLMAP [76] 对不一致的特征较为敏感，容易无法从生成的新视角中提取位姿，我们改用 DUST3R [19] 进行更鲁棒的位姿估计。

对比基线 (Comparison baselines)。 As a diffusion-based generalizable novel view synthesis framework, we compare our method with three diffusion-based baselines: ZeroNVS [12], MotionCtrl [13] and LucidDreamer [14]. Specifically, ZeroNVS [12] is finetuned from Zero-1-to-3 [11] and can generate novel views conditioned on a reference image and the relative camera pose. The camera pose is processed as CLIP [62] text embedding and injected into the diffusion U-Net via cross-attention. MotionCtrl [13] is a camera-conditioned video diffusion model finetuned from SVD [17]. It can generate consistent novel views from the conditioned reference image and the relative camera pose sequences, which are also processed as high-level embedding and injected into the video diffusion U-Net through cross-attention. LucidDreamer [14] utilizes depth-based warping to synthesize novel views, and employs a pretrained diffusion-based inpainting model [43] to inpaint missing regions in the novel views. For the zero-shot novel view synthesis comparison, we use a single reference image as input for all baselines and our method, since the baselines are only capable of performing single-view novel view synthesis.

作为一个基于扩散、可泛化的新视角合成框架，我们将本方法与三个基于扩散的基线进行比较：ZeroNVS [12]、

MotionCtrl [13] 和 LucidDreamer [14]。具体来说，ZeroNVS [12] 从 Zero-1-to-3 [11] 微调而来，能够以参考图像和相对相机位姿为条件生成新视角；相机位姿被处理为 CLIP [62] 文本嵌入，并通过交叉注意力注入扩散 U-Net。MotionCtrl [13] 是从 SVD [17] 微调而来的相机条件视频扩散模型，能够根据条件参考图像和相对相机位姿序列生成一致的新视角；这些位姿序列同样被处理为高层嵌入，通过交叉注意力注入视频扩散 U-Net。LucidDreamer [14] 利用基于深度的图像变形来合成新视角，并采用预训练的基于扩散的修补模型 [43] 来修补新视角中的缺失区域。在零样本新视角合成对比中，由于这些基线只能进行单视角新视角合成，我们对所有基线和本方法都使用单张参考图像作为输入。

定性对比 (Qualitative comparison)。 The qualitative results are presented in Fig. 3, where the reference images are displayed in the left-most column, and the ground truth novel views are located in the right-most column. The results of LucidDreamer [14] exhibit severe artifacts, since it uses depth-based warping for generating novel views, which is particularly problematic when handling in-the-wild images with unknown camera intrinsic, leading to inaccurate novel views. Moreover, it employs an off-the-shelf inpainting model [43] to refine the warped results, which tends to introduce inconsistencies between the original and inpainted content. Novel views generated by ZeroNVS [12] also exhibit relatively low quality and poor accuracy; the primary reason is that ZeroNVS introduces the camera pose condition into diffusion models through text embedding, which fails to provide precise control over the generation of novel views, leading to sub-optimal results. Similarly, although MotionCtrl [13] can produce novel views with better fidelity, it falls short in generating novel views that precisely align with the given camera conditions. This is because MotionCtrl also adopts a high-level camera embedding to control camera pose, leading to less accurate novel view synthesis. In comparison, our method incorporates explicit point cloud prior and video diffusion model, the results demonstrate the superiority of our method in terms of both pose control accuracy and the overall quality of the generated novel views.

定性结果如图 3 所示，其中参考图像位于最左列，真值新视角位于最右列。LucidDreamer [14] 的结果出现严重伪影，因为它用基于深度的图像变形来生成新视角，在处理相机内参未知的野外图像时尤其容易出问题，导致新视角不准确；此外，它用一个现成的修补模型 [43] 来修饰变形结果，往往在原始内容与修补内容之间引入不一致。ZeroNVS [12] 生成的新视角质量和精度也相对较低，主要原因是它通过文本嵌入把相机位姿条件引入扩散模型，无法对新视角的生成实现精确控制，从而得到次优结果。类似地，MotionCtrl [13] 虽能生成保真度更好的新视角，却难以生成与给定相机条件精确对齐的新视角，这是因为它同样采用高层相机嵌入来控制相机



图 4. Tanks-and-Temples [21] 数据集上场景重建的定性对比。我们用每个场景的 2 张真值训练图像进行训练，并渲染新视角与真值新视角（最右列）比较。Qualitative comparison of scene reconstruction on Tanks-and-Temples [21] dataset. We train each scene using 2 ground truth training images, and render novel views to compare with the ground truth novel view (right-most column).

位姿，使新视角合成的精度下降。相比之下，本方法融合了显式点云先验与视频扩散模型，结果表明本方法在位姿控制精度和生成新视角的整体质量上都更胜一筹。

定量对比 (Quantitative comparison). The quantitative comparison results are reported in Table. 1. Since ZeroNVS [12] and LucidDreamer [14] can only handle squared images, we crop the generated novel views of our method and MotionCtrl [13] to align with ZeroNVS and LucidDreamer when computing the quantitative metrics. In terms of image quality, it can be observed that our approach consistently outperforms the baselines in all the metrics. Specifically, the higher PSNR and SSIM values indicate that our method maintains better image quality and similarity to the ground truth. The lower LPIPS score further demonstrates that our approach generates more perceptually accurate images, while the significantly improved FID score suggests that our method captures the underlying distribution of the data more effectively. In terms of pose accuracy, the reduced R_{dist} and T_{dist} demonstrate the effectiveness of our model design, which enables more accurate pose control in novel view synthesis.

定量对比结果见表 1。由于 ZeroNVS [12] 和 LucidDreamer [14] 只能处理正方形图像，我们在计算定量指标时，把本方法和 MotionCtrl [13] 生成的新视角裁剪至与 ZeroNVS、LucidDreamer 对齐。在图像质量方面，可以看到本方法在所有指标上都稳定优于基线：更高的 PSNR 和 SSIM

表明本方法保持了更好的图像质量以及和真值的相似度；更低的 LPIPS 进一步说明本方法生成的图像在感知上更准确；而 FID 大幅改善则表明本方法更有效地捕捉了数据的底层分布。在位姿精度方面，更小的 R_{dist} 和 T_{dist} 证明了本模型设计的有效性，使新视角合成中的位姿控制更加精确。

4.3 场景重建对比 (Scene Reconstruction Comparison)

数据集与评价指标 (Datasets and evaluation metrics)。

In the scene reconstruction comparison, we use 6 scenes from the Tanks-and-Temples dataset [21] for evaluation. We create a challenging sparse-view benchmark that contains only 2 ground truth training images for each scene, and use 12 views for evaluation. We employ PSNR, SSIM [73], and LPIPS [74] as the evaluation metrics for image quality assessment.

在场景重建对比中，我们使用 Tanks-and-Temples 数据集 [21] 中的 6 个场景进行评测。我们构建了一个具有挑战性的稀疏视角基准，每个场景仅含 2 张真值训练图像，并用 12 个视角进行评测。我们采用 PSNR、SSIM [73] 和 LPIPS [74] 作为图像质量评估指标。

对比基线 (Comparison baselines). We compare our method with three 3D-GS representation-based sparse view reconstruction methods: DNGaussian [77], FSGS [78] and InstantSplat [79]. Specifically, DNGaussian [12] and FSGS [78] utilize point cloud produced by COLAMP [76] as initialization, and leverage

表 2

Tanks-and-Temples [21] 上场景重建的定量对比。我们对每个场景使用 2 张真值训练图像, 并采用 12 个视角进行评测。**Quantitative comparison of scene reconstruction on Tanks-and-Temples [21].** We use 2 ground truth training images for each scene, and adopt 12 views for evaluation.

方法	LPIPS ↓	PSNR ↑	SSIM ↑
DNGaussian [77]	0.331	15.47	0.541
FSGS [78]	0.364	17.53	0.558
InstantSplat [79]	0.275	18.61	0.614
ViewCrafter (本文)	0.245	21.50	0.692

both image supervision and depth regularization for sparse view reconstruction. InstantSplat [79] explores utilizing point cloud produced by DUST3R [19] as initialization, which enables efficient 3D-GS training from sparse images.

我们将本方法与三种基于 3D-GS 表示的稀疏视角重建方法进行比较: DNGaussian [77]、FSGS [78] 和 InstantSplat [79]。具体而言, DNGaussian [12] 和 FSGS [78] 采用 COLMAP [76] 生成的点云作为初始化, 并同时利用图像监督和深度正则化进行稀疏视角重建。InstantSplat [79] 则探索采用 DUST3R [19] 生成的点云作为初始化, 从而能够从稀疏图像高效训练 3D-GS。

定性对比 (Qualitative comparison)。 The qualitative comparison results are presented in Fig. 4. It can be observed the results from DNGaussian [77] exhibit significant artifacts. Similarly, results from FSGS [78] show artifacts when viewed from novel views that deviate from the ground truth training images. Although InstantSplat [79] utilizes DUST3R [19] for point cloud initialization, which better preserves details from the ground truth training images, it fails to recover occlusion regions due to its omission of the densification process [2], resulting in severe holes under novel views. In comparison, our method leverages the priors from video diffusion models, enabling the generation of high-fidelity novel views given only 2 ground truth training images.

定性对比结果如图 4 所示。可以看到, DNGaussian [77] 的结果出现明显伪影。类似地, FSGS [78] 的结果在偏离真值训练图像的新视角下也出现伪影。InstantSplat [79] 虽用 DUST3R [19] 做点云初始化、能较好保留真值训练图像的细节, 但由于省略了稠密化过程 [2], 无法恢复遮挡区域, 导致新视角下出现严重空洞。相比之下, 本方法借助视频扩散模型的先验, 仅凭 2 张真值训练图像即可生成高保真新视角。

定量对比 (Quantitative comparison)。 The quantitative comparison results are presented in Table. 2. It can be observed that our approach consistently outperforms the comparison baselines in all the metrics, further validating the effectiveness

表 3

位姿条件策略的消融研究。除条件信号外, Plücker 模型的架构和训练策略与 ViewCrafter 完全一致。**Ablation study of the pose condition strategy.** Except for the condition signal, the architecture and training strategy of the Plücker model are identical to those of ViewCrafter.

方法	LPIPS ↓	PSNR ↑	SSIM ↑	FID ↓	$R_{\text{dist}} \downarrow$	$T_{\text{dist}} \downarrow$
Plücker 模型	0.370	17.51	0.546	49.33	2.688	2.570
本文方法	0.270	20.25	0.649	38.17	0.552	0.983

表 4

不同训练范式的定量对比。我们分析同时训练视频去噪 U-Net 的空间层与时序层的有效性, 以及渐进式训练策略和推理更多帧带来的收益。**Quantitative comparison of different training paradigms.** We analyze the effectiveness of training both the spatial and temporal layers of the video denoising U-Net, as well as the benefits of the progressive training strategy and inference with more frames.

训练范式	LPIPS ↓	PSNR ↑	SSIM ↑	FID ↓
仅训练空间层	0.301	18.82	0.595	42.30
直接在 576×1024 上训练	0.314	18.55	0.582	41.01
16 帧模型	0.289	19.07	0.610	38.43
本文方法	0.280	19.52	0.615	37.77

of our method in scene reconstruction from sparse views.

定量对比结果见表 2。可以看到, 本方法在所有指标上都稳定优于对比基线, 进一步印证了本方法在稀疏视角场景重建中的有效性。

4.4 消融研究 (Ablation Study)

位姿条件策略的讨论 (Discussion on pose condition strategy)。In our method, we utilize point cloud renders as an explicit condition for the video diffusion model, enabling highly accurate pose control for novel view synthesis. Some concurrent works [57], [80] adopt Plücker coordinates [81] as pose condition for pose-controllable video generation. To compare the pose accuracy between our point cloud-based pose condition strategy and the Plücker coordinate-based pose condition strategy, we train a Plücker coordinate-conditioned video diffusion model (Plücker model for short) that accepts Plücker coordinates as conditions for synthesizing novel views. The Plücker coordinate describes per-pixel motion; For a given RGB frame and its camera pose, its Plücker coordinate shares the same spatial size with the RGB frame and comprises 6 channels for each pixel location. Given a pose sequence, we resize the corresponding Plücker coordinates to align with the size of the latent space, and concatenate them with noise along the channel dimension. Except for the pose condition strategy, we maintain the rest

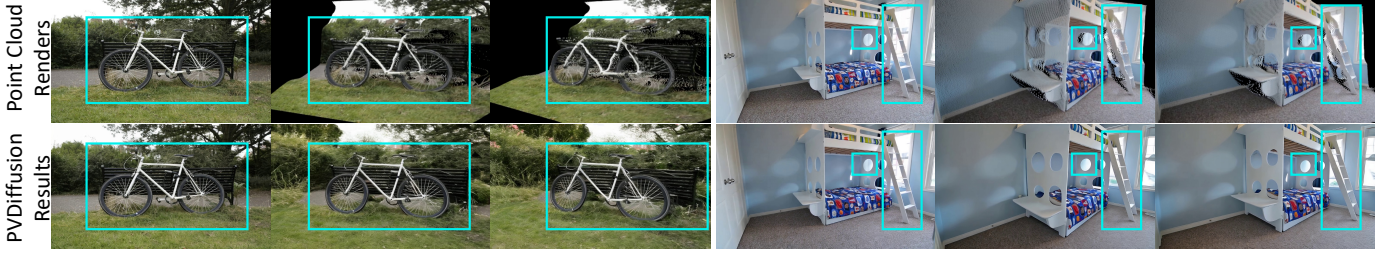


图 5. 点云条件的鲁棒性。上行展示点云渲染结果，下行展示 ViewCrafter 相应生成的新视角（放大观看效果最佳）。**Robustness for point cloud condition.** We show the point cloud render results and the corresponding novel views generated by ViewCrafter in the top and bottom rows, respectively (Best viewed with zoom-in).



图 6. 位姿条件策略的定性消融。除条件信号外，我们让 Plücker 模型的架构和训练策略与 ViewCrafter 完全一致。**Qualitative ablation of pose condition strategy.** We make the architecture and training strategy of the Plücker model to be identical to those of ViewCrafter, with the exception of the condition signal.

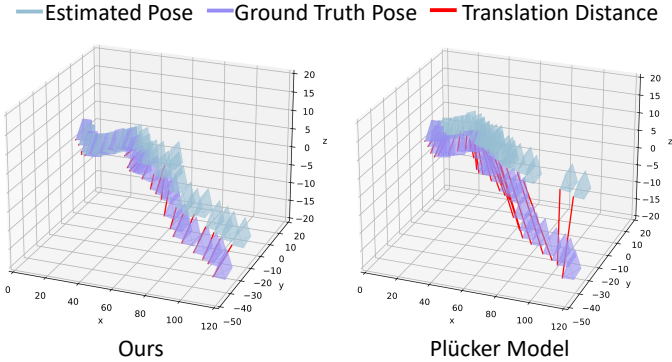


图 7. 位姿精度可视化。我们比较真值相机位姿与从 ViewCrafter、Plücker 模型生成的新视角所估计位姿之间的对齐程度。**Visualization of pose accuracy.** We compare the alignment level between the ground truth camera poses and the poses estimated from the generated novel views of ViewCrafter and the Plücker model.

of the architecture of the Plücker model to be identical to ViewCrafter, and train the model at 320×512 resolution, following the training details reported in Section 4.1. We conduct a zero-shot novel view synthesis comparison between ViewCrafter (320×512 resolution) and the Plücker model. The qualitative and quantitative results are shown in Fig. 6 and Table. 3, which demonstrates that the point cloud-based pose condition strategy employed in ViewCrafter achieves more accurate pose control in novel view synthesis. We also observed that the

Plücker model prone to ignore the high-frequency movements of the camera. Fig. 7 presents an example, where we compare the alignment level between the ground truth camera poses and the poses estimated from the generated novel views. The results show that the poses estimated from the novel views generated by ViewCrafter align more closely with the ground truth poses, further demonstrating the effectiveness of our point cloud-based pose condition strategy.

在本方法中，我们利用点云渲染作为视频扩散模型的显式条件，从而在新视角合成中实现高度精确的位姿控制。一些同期工作 [57], [80] 采用普吕克坐标 [81] 作为位姿条件来进行可控位姿的视频生成。为比较我们基于点云的位姿条件策略与基于普吕克坐标的位姿条件策略在位姿精度上的差异，我们训练了一个以普吕克坐标为条件的视频扩散模型（简称 Plücker 模型），它接受普吕克坐标作为条件来合成新视角。普吕克坐标描述逐像素的运动：对给定的一帧 RGB 图像及其相机位姿，其普吕克坐标与该 RGB 帧具有相同的空间尺寸，并在每个像素位置包含 6 个通道。给定一位姿序列，我们把对应的普吕克坐标缩放到与潜在空间尺寸对齐，并沿通道维与噪声拼接。除位姿条件策略外，我们让 Plücker 模型的其余架构与 ViewCrafter 完全一致，并按第 4.1 节所述的训练细节，在 320×512 分辨率下训练该模型。我们在 ViewCrafter (320×512 分辨率) 与 Plücker 模型之间进行零样本新视角合成对比。定性与定量结果分别见图 6 和表 3，表

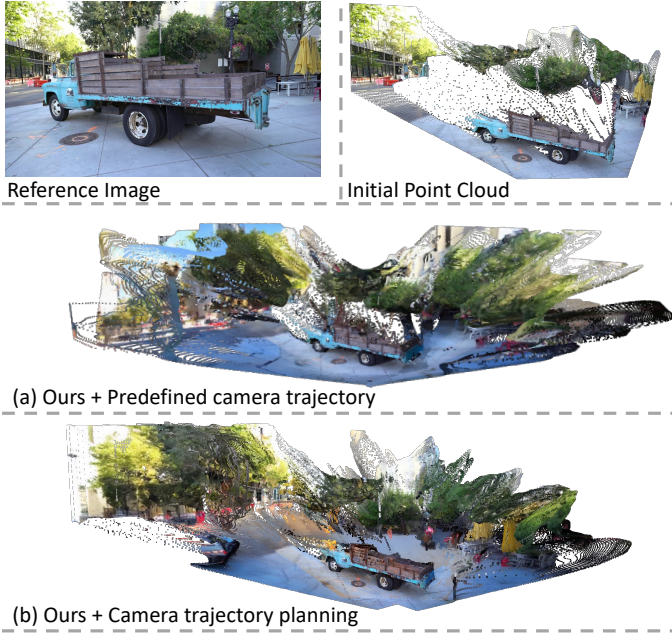


图 8. 相机轨迹规划的消融研究。用预定义相机轨迹重建的点云无法有效补全遮挡区域；相比之下，通过我们的相机轨迹规划算法生成的点云能更有效地揭示场景的遮挡区域，提升点云的整体重建质量。

Ablation study on the camera trajectory planning. The point cloud reconstructed using the predefined camera trajectory fails to effectively complete the occlusion region. In contrast, the point cloud generated through our camera trajectory planning algorithm reveals the occlusion region of the scene more effectively, enhancing the overall reconstruction quality of the point cloud.

明 ViewCrafter 采用的基于点云的位姿条件策略在新视角合成中实现了更精确的位姿控制。我们还观察到，Plücker 模型容易忽略相机的高频运动。图 7 给出一个例子，其中我们比较真值相机位姿与从生成新视角估计出的位姿之间的对齐程度。结果显示，从 ViewCrafter 生成的新视角估计出的位姿与真值位姿对齐得更紧密，进一步证明了我们基于点云的位姿条件策略的有效性。

点云条件的鲁棒性 (Robustness for point cloud condition). ViewCrafter utilizes point cloud render results as conditions, enabling highly accurate pose control. However, these renders may contain artifacts and geometric distortions. Fig. 5 presents an example: the first row shows that the conditioned point cloud renders exhibit occlusions and missing regions, as well as geometric distortions along the boundary of the foreground. The second row displays the corresponding novel views produced by ViewCrafter, demonstrating its ability to fill in the holes and correct the inaccurate geometry. This demonstrate that ViewCrafter has developed a comprehensive understanding of the 3D world, allowing it to generate high-quality novel views from imperfect conditional information and exhibit robustness to the point cloud conditions.

ViewCrafter 以点云渲染结果作为条件，从而实现高度精确的位姿控制。然而，这些渲染可能含有伪影和几何畸变。图 5 给出一个例子：第一行显示，作为条件的点云渲染出现遮挡和缺失区域，以及前景边界处的几何畸变；第二行展示 ViewCrafter 相应生成的新视角，表明它能够填补空洞并纠正不准确的几何。这说明 ViewCrafter 已形成对三维世界的全面理解，使其能够从不完美的条件信息生成高质量新视角，并对点云条件表现出鲁棒性。

训练范式的消融 (Ablation on training paradigm). We examine the effectiveness of the adopted training paradigm in this ablation study. To evaluate the choice of training which module of the video denoising U-Net, we compare the novel view synthesis quality of training only the spatial layers and training both the spatial and temporal layers (Ours), the results are shown in the first row of Table. 4. To assess the importance of progressive training, we provide a comparison between directly training the model at 576×1024 resolution and training the model using the progressive training strategy (Ours), the results are reported in the second row of Table. 4. Additionally, we have observed that the inference length of ViewCrafter influences the quality of novel view synthesis. Specifically, for the same range of view change, inference with more frames improves the temporal consistency of the generated frames. To balance the computational cost and synthesis quality, we train two models: a base model that infers 16 frames and a stronger model (Ours) that infers 25 frames. We present a comparison between the two models in the third row of Table. 4. The results above showcase the effectiveness of the implemented training paradigm.

本消融研究考察所采用训练范式的有效性。为评估应训练视频去噪 U-Net 中哪个模块的选择，我们比较仅训练空间层与同时训练空间层和时序层（本文方法）的新视角合成质量，结果见表 4 第一行。为评估渐进式训练的重要性，我们比较直接在 576×1024 分辨率下训练模型与采用渐进式训练策略训练模型（本文方法），结果见表 4 第二行。此外，我们观察到 ViewCrafter 的推理长度会影响新视角合成质量：在相同的视角变化范围内，推理更多帧能提升生成帧的时序一致性。为在计算开销与合成质量之间取得平衡，我们训练了两个模型：推理 16 帧的基础模型与推理 25 帧的更强模型（本文方法）。两个模型的比较见表 4 第三行。上述结果展示了所实现训练范式的有效性。

相机轨迹规划的消融 (Ablation on camera trajectory planning). In this ablation study, we assess the effectiveness of our proposed camera trajectory planning algorithm for revealing occlusions and completing point clouds. An example is presented in Fig. 8, where we compare the point cloud generated by iterative view synthesis using a predefined circular camera trajectory (similar as [14]) with that produced using our camera trajectory planning algorithm. Given a reference image and

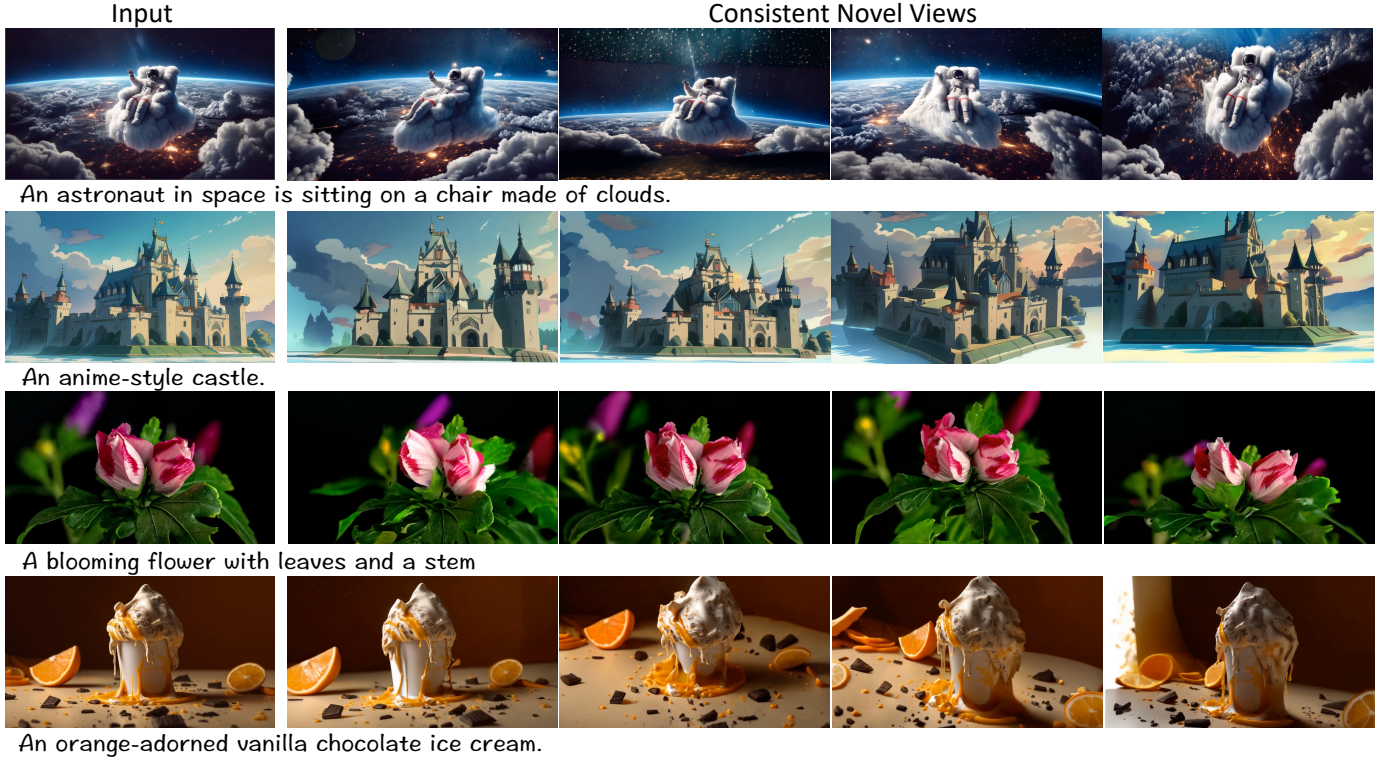


图 9. **文本到 3D 生成结果**。最左列展示输入文本提示以及 T2I 模型生成的对应图像，其余各列展示 ViewCrafter 基于该生成图像所产生的一致新视角。**Text-to-3D generation results.** The leftmost column displays the input text prompt and the corresponding image generated by T2I model, while the subsequent columns show the consistent novel views produced by our ViewCrafter based on the generated image.

an initial point cloud, we adopt a quarter-sphere searching space centered at the origin of the initial point cloud’s world coordinate system, with the radius set to the depth of the center pixel of the reference image. We begin with exploring the left half area of the searching space, with the parameters for camera trajectory planning set to $N = 3$, $K = 5$, and $\Theta = 0.6$. Accordingly, the circular camera trajectory is set as uniformly moving the camera three times from the reference pose to the left direction of the searching space, with each movement measuring 20 degrees. Subsequently, we apply the same parameters to explore the right half of the searching space. The final generated point cloud are shown in Fig. 8(a) and Fig. 8(b). In Fig. 8(a), It can be observed that the point cloud reconstructed using the predefined circular camera trajectory results in ineffective completion of the occlusion region. In comparison, Fig. 8(b) presents the reconstruction using our camera trajectory planning algorithm. The more complete reconstruction results demonstrate that it can effectively reveal occlusion regions of the scene, improving the overall scene reconstruction quality.

本消融研究评估我们提出的相机轨迹规划算法在揭示遮挡和补全点云方面的有效性。图 8 给出一个例子，其中我们比较两种做法生成的点云：一种用预定义的圆形相机轨迹（与

[14] 类似）进行迭代式视角合成，另一种用我们的相机轨迹规划算法。给定一张参考图像和一个初始点云，我们采用一个以初始点云世界坐标系原点为中心的四分之一球面搜索空间，半径设为参考图像中心像素的深度。我们先探索搜索空间的左半区，相机轨迹规划参数设为 $N = 3$ 、 $K = 5$ 、 $\Theta = 0.6$ 。相应地，圆形相机轨迹设置为：从参考位姿出发，向搜索空间的左方向均匀移动相机三次，每次移动 20 度。随后，我们用相同参数探索搜索空间的右半区。最终生成的点云如图 8(a) 和图 8(b) 所示。在图 8(a) 中可以看到，用预定义圆形相机轨迹重建出的点云未能有效补全遮挡区域；相比之下，图 8(b) 展示了用我们的相机轨迹规划算法进行的重建。更完整的重建结果表明，该算法能够有效揭示场景的遮挡区域，提升整体场景重建质量。

4.5 文本到 3D 生成 (Text-to-3D Generation)

In addition to synthesizing novel views of real-world images, we also explore the application of combining our framework with creative text-to-image (T2I) diffusion models for text-to-3D generation. To accomplish this, given a text prompt, we first adopt T2I models to generate a corresponding image, then apply ViewCrafter to synthesize consistent novel views. The results are shown in Fig. 9.

除对真实世界图像合成新视角外，我们还探索将本框架与富有创意的文本到图像 (T2I) 扩散模型结合，用于文本到 3D 生成。其做法是：给定一段文本提示，我们先用 T2I 模型生成对应图像，再用 ViewCrafter 合成一致的新视角。结果如图 9 所示。

5 结论与局限 (Conclusion and limitation)

This work presents ViewCrafter, a novel view synthesis framework that combines video diffusion models and point cloud priors for high-fidelity and accurate novel view synthesis. Our method overcomes the limitations of existing approaches by providing generalization ability for various scene types and adaptability for both single and sparse image inputs, while maintaining consistency and accuracy in the quality of novel views. Additionally, we introduce an iterative view synthesis method and an adaptive camera trajectory planning procedure that facilitate long-range novel view synthesis and automatic camera trajectory generation for diverse scenes. Beyond novel view synthesis, we explore the efficient optimization of a 3D-GS representation for real-time, high frame-rate novel view rendering, and adapting our framework for text-to-3D generation.

本文提出 ViewCrafter，一个结合视频扩散模型与点云先验、用于高保真且精确的新视角合成框架。本方法克服了现有方法的局限：既能泛化到各类场景，又能适配单张和稀疏图像两种输入，同时在新视角质量上保持一致性与精确性。此外，我们提出一种迭代式视角合成方法和一套自适应相机轨迹规划流程，支撑长距离新视角合成，并为各类场景自动生成相机轨迹。除新视角合成外，我们还探索了高效优化 3D-GS 表示以实现实时、高帧率的新视角渲染，以及将本框架适配到文本到 3D 生成。

局限 (Limitations). Despite its advantages, our method still has several limitations. Firstly, it may encounter challenges in synthesizing novel views with a very large view range given limited 3D clues, such as generating a front-view image from only a back-view image. Additionally, we leverage point clouds as an explicit prior and have validated the robustness of our method for low-quality point clouds. However, challenges may still persist in scenes where the conditioned point clouds are significantly inaccurate. Furthermore, as a video diffusion model, our method necessitates multi-step denoising during the inference process, which requires a relatively higher computing cost.

尽管有诸多优势，本方法仍存在若干局限。其一，在三维线索有限的情况下合成视角范围极大的新视角时，本方法可能遇到困难，例如仅凭一张后视图生成前视图。其二，我们以点云作为显式先验，并验证了本方法对低质量点云的鲁棒性；但在条件点云严重不准确的场景中，困难可能依然存在。

其三，作为一个视频扩散模型，本方法在推理过程中需要多步去噪，计算开销相对较高。

参考文献

- [1] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *ECCV*, 2020.
- [2] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM TOG*, 2023.
- [3] O. Wiles, G. Gkioxari, R. Szeliski, and J. Johnson, "Synsin: End-to-end view synthesis from a single image," in *CVPR*, 2020.
- [4] R. Rombach, P. Esser, and B. Ommer, "Geometry-free view synthesis: Transformers and no 3d priors," in *ICCV*, 2021.
- [5] C. Rockwell, D. F. Fouhey, and J. Johnson, "Pixelsynth: Generating a 3d-consistent experience from a single image," in *ICCV*, 2021.
- [6] B. Park, H. Go, and C. Kim, "Bridging implicit and explicit geometric transformation for single-image view synthesis," *IEEE TPAMI*, 2024.
- [7] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely, "Stereo magnification: Learning view synthesis using multiplane images," *ACM TOG*, 2018.
- [8] Y. Han, R. Wang, and J. Yang, "Single-view view synthesis in the wild with learned adaptive multiplane images," in *SIGGRAPH Conference*, 2022.
- [9] R. Tucker and N. Snavely, "Single-view view synthesis with multiplane images," in *CVPR*, 2020.
- [10] A. Yu, V. Ye, M. Tancik, and A. Kanazawa, "pixelnerf: Neural radiance fields from one or few images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4578–4587.
- [11] R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick, "Zero-1-to-3: Zero-shot one image to 3d object," in *ICCV*, 2023.
- [12] K. Sargent, Z. Li, T. Shah, C. Herrmann, H.-X. Yu, Y. Zhang, E. R. Chan, D. Lagun, L. Fei-Fei, D. Sun, and J. Wu, "ZeroNVS: Zero-shot 360-degree view synthesis from a single real image," in *CVPR*, 2024.
- [13] Z. Wang, Z. Yuan, X. Wang, T. Chen, M. Xia, P. Luo, and Y. Shan, "Motionctrl: A unified and flexible motion controller for video generation," in *SIGGRAPH Conference*, 2024.
- [14] J. Chung, S. Lee, H. Nam, J. Lee, and K. M. Lee, "Luciddreamer: Domain-free generation of 3d gaussian splatting scenes," *arXiv preprint arXiv:2311.13384*, 2023.
- [15] J. Shriram, A. Trevithick, L. Liu, and R. Ramamoorthi, "Realdreamer: Text-driven 3d scene generation with inpainting and depth diffusion," *arXiv preprint arXiv:2404.07199*, 2024.
- [16] P.-Y. Lab and T. A. etc., "Open-sora-plan," <https://github.com/PKU-YuanGroup/Open-Sora-Plan>, 2024. [Online]. Available: <https://doi.org/10.5281/zenodo.10948109>
- [17] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts *et al.*, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv preprint arXiv:2311.15127*, 2023.

- [18] J. Xing, M. Xia, Y. Zhang, H. Chen, X. Wang, T.-T. Wong, and Y. Shan, "Dynamicrafter: Animating open-domain images with video diffusion priors," *arXiv preprint arXiv:2310.12190*, 2023.
- [19] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "Dust3r: Geometric 3d vision made easy," in *CVPR*, 2024.
- [20] V. Leroy, Y. Cabon, and J. Revaud, "Grounding image matching in 3d with mast3r," *arXiv:2406.09756*, 2024.
- [21] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM TOG*, 2017.
- [22] J. Reizenstein, R. Shapovalov, P. Henzler, L. Sbordone, P. Labatut, and D. Novotny, "Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction," in *ICCV*, 2021.
- [23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.
- [24] A. Treuithick, M. Chan, M. Stengel, E. Chan, C. Liu, Z. Yu, S. Khamis, M. Chandraker, R. Ramamoorthi, and K. Nagano, "Real-time radiance fields for single-image portrait view synthesis," *ACM TOG*, 2023.
- [25] W. Yu, Y. Fan, Y. Zhang, X. Wang, F. Yin, Y. Bai, Y.-P. Cao, Y. Shan, Y. Wu, Z. Sun *et al.*, "Nofa: Nerf-based one-shot facial avatar reconstruction," in *SIGGRAPH Conference*, 2023.
- [26] Y. Hong, K. Zhang, J. Gu, S. Bi, Y. Zhou, D. Liu, F. Liu, K. Sunkavalli, T. Bui, and H. Tan, "Lrm: Large reconstruction model for single image to 3d," in *ICLR*, 2024.
- [27] D. Charatan, S. L. Li, A. Tagliasacchi, and V. Sitzmann, "pixel-splat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction," in *CVPR*, 2024.
- [28] Y. Chen, H. Xu, C. Zheng, B. Zhuang, M. Pollefeys, A. Geiger, T.-J. Cham, and J. Cai, "Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images," in *ECCV*, 2024.
- [29] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, 2020.
- [30] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *ICLR*, 2021.
- [31] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.
- [32] W. Yu, L. Yuan, Y.-P. Cao, X. Gao, X. Li, L. Quan, Y. Shan, and Y. Tian, "Hifi-123: Towards high-fidelity one image to 3d content generation," in *ECCV*, 2024.
- [33] J. Tang, T. Wang, B. Zhang, T. Zhang, R. Yi, L. Ma, and D. Chen, "Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior," in *ICCV*, 2023.
- [34] J. Sun, B. Zhang, R. Shao, L. Wang, W. Liu, Z. Xie, and Y. Liu, "Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior," in *ICLR*, 2024.
- [35] E. R. Chan, K. Nagano, M. A. Chan, A. W. Bergman, J. J. Park, A. Levy, M. Aittala, S. De Mello, T. Karras, and G. Wetzstein, "Generative novel view synthesis with 3d-aware diffusion models," in *ICCV*, 2023.
- [36] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsani, A. Kembhavi, and A. Farhadi, "Objaverse: A universe of annotated 3d objects," in *CVPR*, 2023.
- [37] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su *et al.*, "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.
- [38] D. Watson, W. Chan, R. M. Brualdi, J. Ho, A. Tagliasacchi, and M. Norouzi, "Novel view synthesis with diffusion models," in *ICLR*, 2023.
- [39] J. Reizenstein, R. Shapovalov, P. Henzler, L. Sbordone, P. Labatut, and D. Novotny, "Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction," in *ICCV*, 2021.
- [40] X. Yu, M. Xu, Y. Zhang, H. Liu, C. Ye, Y. Wu, Z. Yan, C. Zhu, Z. Xiong, T. Liang *et al.*, "Mvingnet: A large-scale dataset of multi-view images," in *CVPR*, 2023.
- [41] R. Wu, B. Mildenhall, P. Henzler, K. Park, R. Gao, D. Watson, P. P. Srinivasan, D. Verbin, J. T. Barron, B. Poole *et al.*, "Reconfusion: 3d reconstruction with diffusion priors," in *CVPR*, 2024.
- [42] J. Zhang, X. Li, Z. Wan, C. Wang, and J. Liao, "Text2nerf: Text-driven 3d scene generation with neural radiance fields," *IEEE TVCG*, 2024.
- [43] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.
- [44] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *ICCV*, 2023.
- [45] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, and Y. Shan, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," in *AAAI*, 2024.
- [46] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee, "Gligen: Open-set grounded text-to-image generation," in *CVPR*, 2023.
- [47] J. Xing, H. Liu, M. Xia, Y. Zhang, X. Wang, Y. Shan, and T.-T. Wong, "Toonrafter: Generative cartoon interpolation," *arXiv preprint arXiv:2405.17933*, 2024.
- [48] J. Xing, M. Xia, Y. Liu, Y. Zhang, Y. Zhang, Y. He, H. Liu, H. Chen, X. Cun, X. Wang *et al.*, "Make-your-video: Customized video generation using textual and structural guidance," *IEEE TVCG*, 2024.
- [49] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, "Structure and content-guided video synthesis with diffusion models," in *ICCV*, 2023.
- [50] S. Yin, C. Wu, J. Liang, J. Shi, H. Li, G. Ming, and N. Duan, "Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory," *arXiv preprint arXiv:2308.08089*, 2023.
- [51] M. Niu, X. Cun, X. Wang, Y. Zhang, Y. Shan, and Y. Zheng, "Mofa-video: Controllable image animation via generative motion field adaptations in frozen image-to-video diffusion model," *arXiv preprint arXiv:2405.20222*, 2024.
- [52] E. Peruzzo, V. Goel, D. Xu, X. Xu, Y. Jiang, Z. Wang, H. Shi, and N. Sebe, "Vase: Object-centric appearance and shape manipulation of real videos," *arXiv preprint arXiv:2401.02473*, 2024.
- [53] Y. Guo, C. Yang, A. Rao, Y. Wang, Y. Qiao, D. Lin, and B. Dai, "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning," *arXiv preprint arXiv:2307.04725*, 2023.
- [54] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang,

- L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” in *ICLR*, 2022.
- [55] N. Müller, K. Schwarz, B. Rössle, L. Porzi, S. R. Bulò, M. Nießner, and P. Kotschieder, “Multidiff: Consistent novel view synthesis from a single image,” in *CVPR*, 2024.
- [56] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, “ScanNet: Richly-annotated 3d reconstructions of indoor scenes,” in *CVPR*, 2017.
- [57] D. Xu, W. Nie, C. Liu, S. Liu, J. Kautz, Z. Wang, and A. Vahdat, “Camco: Camera-controllable 3d-consistent image-to-video generation,” *arXiv preprint arXiv:2406.02509*, 2024.
- [58] H. He, Y. Xu, Y. Guo, G. Wetzstein, B. Dai, H. Li, and C. Yang, “Cameractrl: Enabling camera control for text-to-video generation,” *arXiv preprint arXiv:2404.02101*, 2024.
- [59] V. Sitzmann, S. Rezchikov, B. Freeman, J. Tenenbaum, and F. Durand, “Light field networks: Neural scene representations with single-evaluation rendering,” in *NeurIPS*, 2021.
- [60] G. Metzger, E. Richardson, O. Patashnik, R. Giryes, and D. Cohen-Or, “Latent-nerf for shape-guided generation of 3d shapes and textures,” in *CVPR*, 2023.
- [61] F. Plastria, “The weiszfeld algorithm: proof, amendments and extensions, ha eiselt and v. marianov (eds.) foundations of location analysis, international series in operations research and management science,” 2011.
- [62] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *ICML*, 2021.
- [63] R. Gao, A. Holynski, P. Henzler, A. Brussee, R. Martin-Brualla, P. Srinivasan, J. T. Barron, and B. Poole, “Cat3d: Create anything in 3d with multi-view diffusion models,” *arXiv preprint arXiv:2405.10314*, 2024.
- [64] R. Zeng, Y. Wen, W. Zhao, and Y.-J. Liu, “View planning in robot active vision: A survey of systems, algorithms, and applications,” *Computational Visual Media*, vol. 6, 2020.
- [65] H. Dhami, V. D. Sharma, and P. Tokekar, “Pred-nbv: Prediction-guided next-best-view planning for 3d object reconstruction,” in *IROS*, 2023.
- [66] L. Jin, X. Chen, J. Rükin, and M. Popović, “Neu-nbv: Next best view planning using uncertainty estimation in image-based neural rendering,” in *IROS*, 2023.
- [67] S. Isler, R. Sabzevari, J. Delmerico, and D. Scaramuzza, “An information gain formulation for active volumetric 3d reconstruction,” in *ICRA*, 2016.
- [68] D. Peralta, J. Casimiro, A. M. Nilles, J. A. Aguilar, R. Atienza, and R. Cajote, “Next-best view policy for 3d reconstruction,” in *ECCVW*, 2020.
- [69] M.-L. Shih, S.-Y. Su, J. Kopf, and J.-B. Huang, “3d photography using context-aware layered depth inpainting,” in *CVPR*, 2020.
- [70] L. Ling, Y. Sheng, Z. Tu, W. Zhao, C. Xin, K. Wan, L. Yu, Q. Guo, Z. Yu, Y. Lu *et al.*, “Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision,” in *CVPR*, 2024.
- [71] N. Ravi, J. Reizenstein, D. Novotny, T. Gordon, W.-Y. Lo, J. Johnson, and G. Gkioxari, “Accelerating 3d deep learning with pytorch3d,” *arXiv:2007.08501*, 2020.
- [72] J. Ho and T. Salimans, “Classifier-free diffusion guidance,” *arXiv preprint arXiv:2207.12598*, 2022.
- [73] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE TIP*, vol. 13, no. 4, pp. 600–612, 2004.
- [74] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *CVPR*, 2018.
- [75] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” in *NeurIPS*, 2017.
- [76] J. L. Schönberger and J.-M. Frahm, “Structure-from-motion revisited,” in *CVPR*, 2016.
- [77] J. Li, J. Zhang, X. Bai, J. Zheng, X. Ning, J. Zhou, and L. Gu, “Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization,” in *CVPR*, 2024.
- [78] Z. Zhu, Z. Fan, Y. Jiang, and Z. Wang, “Fsgs: Real-time few-shot view synthesis using gaussian splatting,” in *ECCV*, 2024.
- [79] Z. Fan, W. Cong, K. Wen, K. Wang, J. Zhang, X. Ding, D. Xu, B. Ivanovic, M. Pavone, G. Pavlakos *et al.*, “Instantsplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds,” *arXiv:2403.20309*, 2024.
- [80] S. Bahmani, I. Skorokhodov, A. Siarohin, W. Menapace, G. Qian, M. Vasilkovsky, H.-Y. Lee, C. Wang, J. Zou, A. Tagliasacchi *et al.*, “Vd3d: Taming large video diffusion transformers for 3d camera control,” *arXiv preprint arXiv:2407.12781*, 2024.
- [81] Y.-B. Jia, “Plücker coordinates for lines in the space,” *Problem Solver Techniques for Applied Computer Science, Com-S-477/577 Course Handout*, 2020.