随着生成建模技术的快速发展，创建各种数字内容的能力得到显著提高，如 2D 图像、视频和 3D 场景。然而，生成连续且高质量的动态 4D 场景仍然具有挑战性。动态 4D 场景通常由动态神经辐射场（NeRF）表示，旨在从任何视角显示一致的外观、几何形状和运动。最近的工作引入了时间集变量，以优化基于 3D 高斯散射（3D GS）技术的高斯变形场，从而产生一种新颖的 4D 表示，称为 4D 高斯散射（4D GS）。这种方法旨在实现实时动态场景渲染的高效训练和存储，同时保持高质量的输出。动态 4D 场景生成通常依赖于 NeRF 的概念，能够从任何角度精确再现物体的外观、几何结构和运动状态。这项技术结合了视频处理和 3D 建模的优势，生成从简单文本和视频描述到复杂 4D 场景的高质量内容。例如，Hexplane [Cao and Johnson(2023)] 技术使得能够直接从文本生成动态视频，然后通过高级扩散模型将其转换为 4D 演示，大大丰富了内容创作的可能性。新兴技术如 CONSISTENT4D [Jiang et al.(2024)] 和 MAV3D [Singer et al.(2023)] ，通过优化 NeRF 框架和扩散模型，尝试从单一视频源生成连贯的 4D 场景。这些方法有效地将视频信息与动态 3D 模型结合，以生成在细节和动态纹理上丰富的 4D 内容。此外，像 Animate124 [Zhao et al.(2023)] 这样的技术探索将单个图像转变为动态 3D 视频，进一步扩展了动态场景生成的应用。尽管这些先进技术具有潜力，但在清晰度和时间效率方面仍然存在挑战。目前的方法可能需要较长处理时间才能生成高质量的动态 NeRF，并且生成的动态场景的清晰度需要改进。在确保质量的同时提高效率和清晰度仍然是当前 4D 生成技术研究的重点。4D GS 方法采用轻量级的 MLP，预测新时间戳的 3D 高斯变形，为动态场景提供了一种新颖的显示表示。这种方法显著减少了优化时间，并实现了实时处理，同时保持高质量的可变形高斯喷溅。尽管 4D 高斯散点（4D GS）为 4D 场景表示提供了一种高效的方法，并且显著改善了 4D 内容的优化时间，但使用 4D GS 技术从单一图像生成 4D 内容仍然面临挑战。这些挑战包括背景点的缺失和摄像机定位不准确的问题，这些都妨碍了动态场景的精确优化。此外，在运动过程中，高斯点的位移可能导致 3D 对象表面撕裂，从而影响渲染的质量和真实性。因此，基于 4D GS 的实时 4D 场景生成的当前方法仍需要进一步优化。为了解决这些挑战，我们提出了一种多视图图像生成模型，该模型捕捉输入对象的全面多视图信息，解决了背景点缺失和摄像机定位不准确等问题，同时显著减少了 4D 场景的生成时间。我们的工作总结如下：1. 我们引入了一个图像矩阵生成模块，该模块从单个输入图像生成多样且时间连贯的多视图图像。该模块提供丰富的空间和时间监督，显著增强了 4D 动态内容的清晰度、一致性和真实感。2. 我们通过直接从生成的图像矩阵初始化和优化 3D 高斯点云来开发紧凑的 3D 场景表示。这种表示通过减少计算开销实现精确重建，为动态建模提供了坚实的基础。3. 我们构建了一个集成框架，通过顺序组合多视图生成、3D 重建和 4D 优化，将单个图像转换为动态 4D 内容。该方法在显著减少生成时间的同时，实现了高视觉保真度和时间一致性。

1 相关工作

使用扩散模型从单张物体图像生成多视图图像方面已经取得了令人振奋的进展。例如，Zero123 利用大规模扩散模型学习的几何先验，在合成数据集上训练条件扩散模型，实现了从单视图输入零样本合成新视角图像。通过将来自不同视角的图像结合为目标图像，Zero123++ 显著克服了 Zero123 生成图像之间无关联的问题，提高了从单张图像生成一致的多视角图像的质量和一致性。尽管从单个输入生成多视图图像方面已经取得了显著进展，但仍然存在一些缺陷。主要问题是在从其他视点生成图像时，难以保持物体的几何一致性。生成的图像可能存在结构和纹理不一致的问题。此外，细节可能丢失的问题也是需要紧迫解决的关键问题，特别是在带有背景的图像中。我们的方法通过使用微调的多视图图像生成模块解决了这些问题。该方法通过最小化视点旋转来增强生成图像的连续性和清晰度。此方法有效地提高了生成图像中物体的几何一致性，同时减少了细节丢失。从图像生成 3D 模型的技术已迅速成熟，出现了许多基于神经辐射场（NeRF）、点云、网格和高斯点散的 3D 模型表示方法。大多数基于多视图图像的方法都专注于创建场景的 3D 表示。One-2-3-45++ 结合了 2D 扩散模型和 3D 原生扩散模型，在多视图条件下，通过一致的多视图图像生成和 3D 重建快速生成 3D 网格。然而，这种基于传统扩散模型的方案通常存在优化时间较长的问题。3D GS 使用 3D 高斯场景表示和实时可微渲染器，实现辐射场的实时渲染，通过优化 3D 高斯属性和密度控制，有效减少了 3D 内容重建所需的时间，并实现高质量的场景表示。尽管尚未解决仅用一张图像输入进行实时 3D 重建的任务，LRM 使用基于 Transformer 的编码器-解码器架构直接从单张图像预测神经辐射场（NeRF）进行 3D 重建，这对于基于单张图像的 3D 重建任务具有重要意义。目前用于 3D 内容生成的方法通常依赖于使用扩散模型对目标模型进行迭代优化。然而，这种方法可能耗时，导致计算成本高。相比之下，基于 Transformer 的编码器-解码器架构提供了更好的可扩展性和效率，但在遮挡区域可能产生模糊纹理，从而导致视觉失真。最近使用 3D 高斯点云（3D GS）进行 3D 内容表示的创新在实时渲染方面表现出色。然而，这些方法在训练期间未观察到的区域可能会出现伪影。我们的方法通过结合一个多视图图像生成模块来解决这些挑战。该模块生成输入图像的额外视图，从而减少由于遮挡区域引起的伪影。此外，我们使用 3D GS 来表示 3D 场景，有效减少计算时间并改善优化效果。动态场景渲染，也称为 4D 生成工作，旨在实现动态 3D 场景的实时渲染，同时具有效率的训练和存储。在这个领域，神经辐射场

(NeRF) 技术被广泛用于表示 4D 场景。例如，Neural-3D-Video 使用时间条件 NeRF 和一系列紧凑的潜码来表示动态场景，并采用分层训练方案和光线重要性采样来显著提高训练速度和生成图像的感知质量。TiNeuVox 结合时间感知体素特征和微坐标变形网络来表示场景，加速了动态辐射场的优化。这项工作在减少训练时间和存储成本的同时，保持了高质量的渲染。然而，动态辐射场重建方法的鲁棒性仍然是一个挑战。RoDynRF 的方法通过从随机捕获的动态单目视频中重建相机轨迹和动态辐射场，实现了专注于鲁棒动态场景的渲染。而 MSTH 则使用哈希码组合来表示动态场景。随着 3D 高斯点绘 (3D GS) 的发展，4D 高斯点绘 (4D GS) 提供了一种创新的方法来渲染动态场景，它通过连续的 4D 表示来处理动态变化。3D GS 函数被操控以适应时间和空间的变化，从而实现实时渲染和高分辨率输出。4D GS 强调计算和存储效率，采用紧凑的网络结构来有效捕捉复杂的动态变化，同时保持高质量的渲染效果。此外，DreamGaussian4D [Ren et al.(2023)] 使用图像到 3D 的框架来拟合静态 3D GS 函数，然后通过学习驱动视频的运动来优化动态表示。最后，4D GS 被导出为一个动画网格序列，并通过视频到视频的过程优化纹理贴图。最近，Stable Video 4D (SV4D) [Xie et al.(2024)] 提供了一种通过多帧多视图一致性实现 4D 生成的动态 3D 内容生成的创新方法。与传统方法依赖于独立训练的视频生成和新颖的视图合成模型不同，SV4D 采用统一的扩散模型，能够从多视图视频生成动态 3D 对象的多视图视频。尽管有这些改进，运动不连续性和高计算量等问题依然存在。我们的方法通过改进图像矩阵模块和高效的 4D GS 优化来增强动态渲染，产生更好的质量和时间连续性。

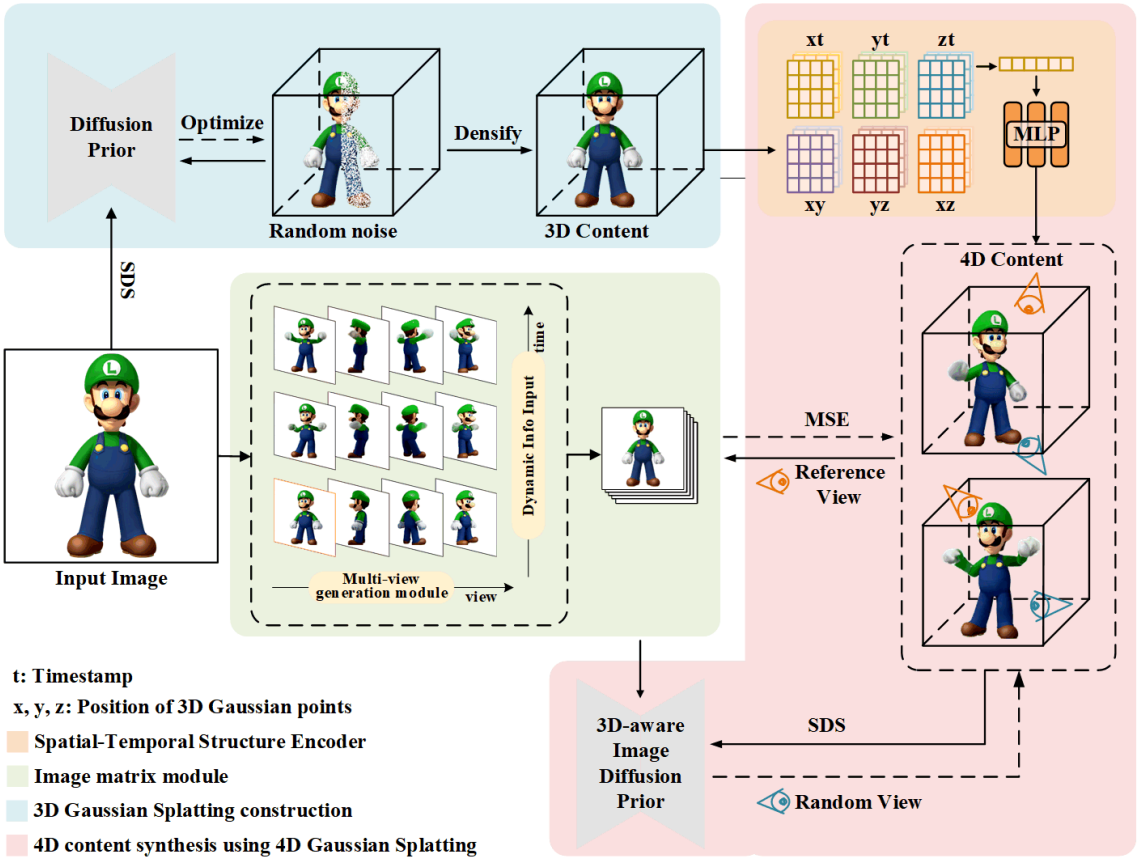

Figure 1: MVG4D 的整体流程。MVG4D 分为三个主要阶段：绿色背景部分是图像矩阵模块，蓝色背景部分是 3D Gaussian Splatting 构建，而红色背景部分则是使用 4D Gaussian Splatting 的 4D 内容合成。

2 方法

从单张图像生成 4D 内容的技术提供了一种创新的 4D 内容生成方法。我们提出的方法 MVG4D 结合了图像矩阵模块与 4D 高斯斑点 (4D GS) 动态内容优化技术。该方法基于 3D 高斯斑点 (3D GS) 技术来优化高斯变形场，这不仅增强了生成 4D 内容的连续性和清晰度，还显著加速了 4D 内容生成过程。我们的研究如图 1 所示，其中概述了三个主要阶段：图像矩阵模块、3D 高斯斑点构建以及使用 4D 高斯斑点进行 4D 内容合成。

2.1 图像矩阵模块

为了从单个输入图像 I_0 构建全面且具时间感知的多视图图像矩阵，我们提出了一个充分利用多视图生成以及引入动态信息的两阶段处理流程。给定输入图像 I_0 ，我们首先利用预训练模块合成一个包含动态信息的视频。该视频随后被分解成一组视频帧 I_t ，其中 t 表示时间戳。虽然这个视频捕捉了时间动态，但它仅包含单视图信息。为了实现多视图动态内容生成，我们微调一个基于扩散的多视图图像生成模型，以合成每个帧 I_t 的新视点。生成的图像矩阵嵌入了时间和视点的多样性，作为优化后续 4D GS 表征的监督信号。我们的方法的核心在于微调一个视图条件的 2D 扩散模型，以便为每个视频帧生成一致的多视图图像。在推理过程中，模型接受原始帧图像和期望的相对相机参数（如角度偏移和深度变化）作为输入，并相应地输出新视图图像。在训练过程中，我们将对象放置在标准的三维坐标系的原点，并模拟一个球形相机设置。相机位置在以对象为中心的球体上，并被限制始终指向原点。设两个相机视点通过球坐标（ θ_1 , ϕ_1 , r_1 ）和（ θ_2 , ϕ_2 , r_2 ）定义，分别代表极角、方位角和半径。它们的相对变换参数化为 $(\Delta\theta, \Delta\phi, \Delta r) = (\theta_2 - \theta_1, \phi_2 - \phi_1, r_2 - r_1)$ 。

扩散模型的训练目标是学习一个函数 f ，使得给定输入图像 x_1 和视点变换 $(\Delta\theta, \Delta\phi, \Delta r)$ 后，模型可以生成一个与从目标视点捕获的真实图像 x_2 非常相似的新视图图像。该过程在图 **??** 中进行了说明。因此，模型学习了一种通用的相机视点控制机制，并且可以在任意相对视点变化下从 x_1 推断目标图像 x_2 。

训练目标表达为：

$$\min_{\theta} \mathbb{E}_{z\sim\mathcal{E}(x),t,\epsilon\sim N(0,1)} \|\epsilon - \epsilon_\theta(z_t, t, c(x, R, T))\|_2^2 \tag{1}$$

在此公式中，x 表示输入图像，$c(x, R, T)$ 代表包含输入图像和目标视点信息的条件嵌入，t 是扩散时间步，$\mathcal{E}$ 是图像编码器，ϵ_θ 是基于 U-Net 的去噪器，而 z_t 则是在时间步 t 时刻的 x 的潜在表示。由于在训练中使用了不同的相对视图参数，每帧 I_t 生成的新颖视图图像可能表现出不一致的视点对齐。这种不一致性在构建的图像矩阵中引入了不匹配，影响了 4D 内容生成的质量。为了解决这个问题，我们进一步使用精心对齐的监督来微调扩散模型，以确保跨视图和时间的连续性。具体来说，在训练过程中，我们最小化生成的新颖视图图像与相应输入帧 I_t 之间的感知差异，强制合成视图与原始输入之间的视觉对齐。这种增强使我们能够生成时空一致的图像矩阵，从而改善 4D GS 的下游优化。

在获得连续且高质量的多视图图像后，我们利用先进的 3D 重建技术将这些 2D 图像转化为 3D GS 模型。这个过程包括两个关键步骤。

我们采用多视图融合技术来整合从各种视角捕获的图像信息。该技术利用图像之间的几何和光照一致性来精确地重建 3D 结构。我们在空间中的随机位置初始化单位尺度、无旋转的三维高斯点云，并在优化期间定期增加其密度。与重建流水线不同，我们一开始使用较少的高斯点云，但更频繁地增加密度以与生成过程保持一致。我们使用评分蒸馏采样（SDS）损失优化三维高斯点云。在每个步骤中，我们采样以物体为中心的随机相机姿态，并从当前视角渲染 RGB 图像。在训练过程中，我们线性降低时间步长 t ，该步长对添加到渲染 RGB 图像的随机噪声进行加权。然后，输入图像作为二维扩散先验，用于使用 SDS 损失优化基础三维高斯点云：

$$\nabla_\Theta \mathcal{L}_{SDS} = \mathbb{E}_{t,p,\epsilon}\left[\omega_{(t)}\left(\epsilon_\phi(I; t, I^r, \triangle p) - \epsilon\right)\frac{\partial I}{\partial \Theta}\right] \tag{2}$$

其中 $\omega_{(t)}$ 是加权函数，ϵ_ϕ 代表通过二维扩散先验 ϕ 预测的噪声，$\triangle p$ 是相对于参考相机 r 的相机姿态相对变化。I 是输入图像，而 I^r 是通过二维扩散模型获得的新姿态图像。通过这种方法，我们可以有效地从二维图像恢复三维几何和纹理信息，为后续生成四维动态模型提供坚实的基础。此阶段的优化目标是增强三维模型的几何精确度和视觉真实感。此外，它旨在确保模型从各个角度查看时的一致性。

2.2 使用四维高斯喷射的四维内容合成

获得包含输入图像的多视角动态信息的图像矩阵和初始化的三维 GS 模型后，我们将静态三维 GS 点云转化为动态四维模型。我们提取每个三维高斯点云的中心坐标 (x, y, z) 和时间戳 t 。然后利用时空结构编码器来计算体素的特征。这些特征通过一个微型 MLP 进行分析和解码，以在时间戳 t 时获得变形的三维高斯点云，渲染变形的点云图像，并匹配第一阶段生成的图像矩阵。计算第一阶段生成的图像矩阵与变形点云渲染的图像之间的均方误差（MSE）损失：

$$\mathcal{L}_{\text{Ref}} = \frac{1}{\mathcal{T}}\sum_{\tau=1}^{\mathcal{T}} \|f(\phi(S, \tau), o_{\text{Ref}}) - I_{\text{Ref}}^\tau\|_2^2 \tag{3}$$

在这个公式中，τ 是一个时间变量，o_{Ref} 表示第一阶段生成的图像矩阵对应的视点信息，相关的图像表示为 I_{Ref}^τ 。使用 4D 变形场模型 f 从该视点渲染图像，并计算渲染图像与 I_{Ref}^τ 之间的误差以

优化模型 f 。为全面解决建模场景被遮挡部分的挑战，我们利用了基于初始阶段生成的多视图图像的 3D 感知图像扩散模型。这种方法使我们能够捕捉更完整和详细的场景信息，然后通过 SDS 损失反向优化 4D 高斯变形场。通过我们对所提出的 MVG4G 方法的研究，我们不仅可以从单个图像中重建精确的 3D 模型，还可以生成动态的 4D 内容。与传统方法相比，我们的技术显著提高了生成效率、动态内容的自然性和视觉连续性。优化后的模型在动态表现中也表现出更高的清晰度和减少的运动闪烁。对于 4D 内容在增强现实和虚拟现实等领域的实际应用，特别是在创造更真实与沉浸式体验方面，这些改进至关重要。

为了评估我们方法在生成高质量动态 4D 内容方面的有效性，我们基于 Objaverse [Deitke et al.(2023)] 等数据集进行了定性和定量评估。对于从单个输入图像生成 4D 内容的任务，设计适当的评价指标对于全面评估生成的多视角和动态内容的质量、一致性和真实性至关重要。我们的方法使用以下指标进行评估：CLIP-I、PSNR、时间、FVD，每个指标针对不同的性能方面：

为了评估从单个图像生成动态 4D 内容的质量，我们将我们的方法 MVG4D 与几个最近的 4D 内容生成方法进行比较，包括 DreamGaussian4D、V4D、4Diffusion 等。我们使用各作者发布的官方代码生成比较结果。

在我们所有的实验中，我们仅使用了一块 NVIDIA RTX 4090 GPU。我们使用一个经过微调的扩散模型处理单个输入图像，以获得图像矩阵监督，从而优化 4D GS 生成动态 4D 内容。

实验结果如表所示，MVG4D 方法生成的 4D 渲染图像与输入图像的 CLIP-I 相似性显著高于现有主流和新颖的 4D 场景生成方法。这表明我们的方法在提高图像相似性方面取得了实质性进展。性能的提升可能归因于我们的多视角图像生成模块生成的多视角图像。这些图像提供额外的空间信息，有效指导了 4D GS 方法的优化。与依赖于视频帧的方法相比，这种创新使我们能够生成更准确的 4D 内容。如表 **??** 所示，与最先进的 4D 场景生成方法相比，MVG4D 生成的 4D 渲染结果显著地获得了更高的 PSNR 分数，突显了所提方法的有效性和先进性。这一优势可能是由于我们使用了 4D 高斯点云来表示 4D 场景，与传统基于 NeRF 的方法相比，更好地积累了空间点信息。如表 **??** 所示，与最先进的 4D 场景生成方法相比，MVG4D 生成的 4D 渲染结果在 FVD-F、FVD-Diag 和 FV4D 方面显著地获得了更低的 FVD 分数，表明生成的动态内容在时间和视图一致性上更好。这一改进主要归因于我们使用了 4D 高斯点云。这主要是由于使用了 3D 高斯散射，它提供了紧凑但富有表现力的初始化，更好地捕捉了空间结构并支持有效的时间变形。

时间是评估 4D 内容生成方法实用性的关键指标，尤其是在 AR/VR 等实时应用中。我们对从单个输入图像到成功生成 4D 内容的总时间进行基准测试。如表 **??** 所示，我们的方法在生成速度上显著优于先前的工作，得益于使用了 4D 高斯散射。与基于 NeRF 的方法相比，我们的轻量级变形网络可以高效地从静态 3D 高斯点云生成时间戳，减少了计算开销。为了评估我们方法在生成高质量动态 4D 内容方面的有效性，我们基于来自 Objaverse 等数据集的五个示例进行了定性评估。如图 **??** 所示，我们提供了基于 4D GS 获得的 4D 场景表示的渲染视图与参考图像的视觉比较。从定性角度来看，生成的动态 4D 场景在不同视点下保持了一致的外观和几何结构，展示了 MVG4D 处理空间变换的卓越能力。此外，生成的场景在纹理细节和运动真实性方面表现优异，从而展示了我们提出的方法在生成高质量动态内容方面的能力。准确的细节渲染和边界处理对于实现高质量动态 4D 内容生成至关重要。如图 **??** 所示，我们比较了基准稳定视频扩散（SVD）方法 [Blattmann et al.(2023)] 与我们提出的 MVG4D 方法的性能。SVD 方法在高频细节上表现出明显的模糊，边缘存在显著的噪声和模糊过渡，降低了清晰度。相比之下，MVG4D 通过保留结构细节并有效抑制噪声实现了更清晰的边界。在边界处理方面，SVD 方法在边缘产生锯齿效应，导致轮廓不平和过渡突兀，而 MVG4D 提供了更平滑和自然的结果，增强了整体视觉保真度。这些结果表明，MVG4D 在细节保留和边界处理上显著优于当前的最新方法，从而带来了更真实且视觉上更引人注目的 4D 动态内容。这种性能提升可以归因于 3D 高斯点云的使用，这比 NeRF 中的基于区间的采样策略提供了更丰富的空间信息，从而实现了更高质量的重建。

我们将本研究中使用的增强扩散模型称为 MVIG（多视图图像生成模块）。图 **??** 展示了一项消融研究的结果，该研究旨在评价视点调整在提高生成的 4D 内容流畅性和清晰度以及增强原始扩散模型方面的有效性。在这项研究中，我们冻结了 3D 内容生成和 4D 内容优化阶段，以隔离后续 4D 表达的影响。我们将原始扩散模型与三个分别针对垂直和水平视角进行微调的模型进行了比较，它们都应用于从相同的输入图像生成动态 4D 内容。我们的微调模型显著优于原始扩散模型，表现出更优越的动作连续性和流畅性，视频帧的闪烁感相比原始模型有所减少。这种改善可归因于多视图图像生成模块，它控制了视图参数的波动，并为 4D GS 模型的优化提供了更一致的视图信息，从而保持了动态 4D 内容的时间一致性。

2.2.1 MVIG 消融实验

表 **??** 展示了我们的多视图图像生成模块 (MVIG) 在提高 4D 内容质量方面的重要性。我们将没有 MVIG 的方法，即直接使用输入的单张图像进行监督的 4D 高斯点集 (4D GS) 优化的方法，称为 “w/o MVIG”。仅使用 MVIG 过程第一阶段的方法标记为 “Half-MVIG”。实验结果表明，完

整的 MVIG，即 “Full-MVIG”，显著增强了所有评估指标。具体来说，“Full-MVIG” 获得了最高的 CLIP-I 和 PSNR 分数，表明它生成了语义一致性最高且高分辨率的 4D 内容。

3 结论

我们提出了 MVG4D，一个从单个静态图像高效生成动态 4D 内容的新框架。我们方法的核心是一个图像矩阵模块，该模块合成了时间上一致且空间上多样的多视图图像集合，为随后的 4D 表示学习提供了密集的监督。通过将这个模块与动态内容优化结合，我们的方法显著提高了渲染 4D 场景的时间连续性、几何一致性和视觉清晰度。实验结果表明，MVG4D 在高保真和现实主义的动态内容重建方面表现出色，与输入图像的外观高度一致。这些进步为动态 4D 内容在增强现实和虚拟现实等领域的应用奠定了坚实的技术基础。未来的工作可能会探索更复杂的场景，包括背景控制和场景合成，以进一步增强我们框架的多样性和表现力。

References

[Blattmann et al.(2023)] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. *CoRR* (2023).

[Cao and Johnson(2023)] Ang Cao and Justin Johnson. 2023. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 130–141.

[Deitke et al.(2023)] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2023. Objaverse: A Universe of Annotated 3D Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13142–13153.

[Jiang et al.(2024)] Yanqin Jiang, Li Zhang, Jin Gao, Weiming Hu, and Yao Yao. 2024. Consistent4D: Consistent 360° Dynamic Object Generation from Monocular Video. In *The Twelfth International Conference on Learning Representations*. `https://openreview.net/forum?id=sPUrdFGepF`

[Ren et al.(2023)] Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. 2023. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142* (2023).

[Singer et al.(2023)] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. 2023. Text-to-4D dynamic scene generation. In *Proceedings of the 40th International Conference on Machine Learning*. 31915–31929.

[Xie et al.(2024)] Yiming Xie, Chun-Han Yao, Vikram Voleti, Huaizu Jiang, and Varun Jampani. 2024. Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. *arXiv preprint arXiv:2407.17470* (2024).

[Zhao et al.(2023)] Yuyang Zhao, Zhiwen Yan, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. 2023. Animate124: Animating one image to 4d dynamic scene. *arXiv preprint arXiv:2311.14603* (2023).