

# You See it, You Got it: Learning 3D Creation on Pose-Free Videos at Scale

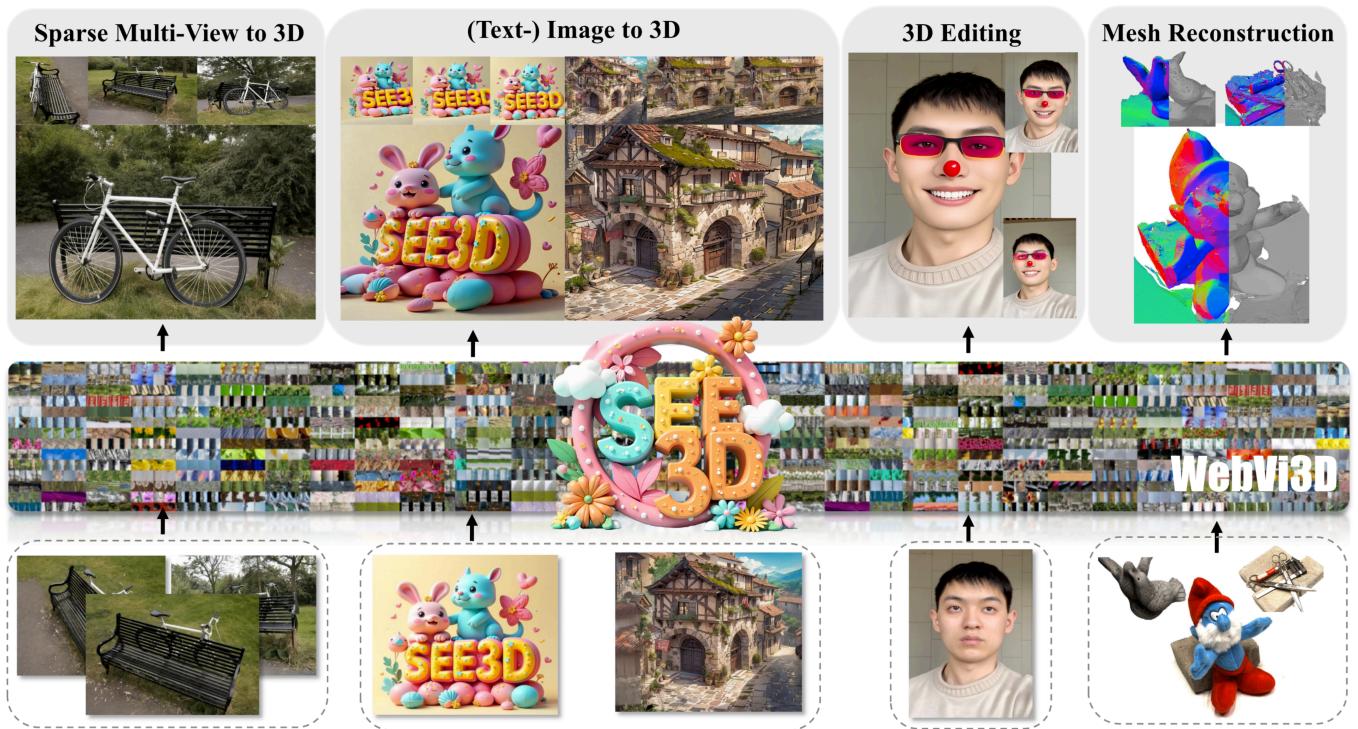


Figure 1. Benefiting from the proposed web-scale dataset WebVi3D, **See3D** enables both object- and scene-level 3D creation, including sparse-view-to-3D, (text-) image-to-3D, and 3D editing. It can also be used for Gaussian Splatting to extract meshes or render images.

## Introduction

近年来，3D生成领域取得了显著进展，在虚拟现实、娱乐和模拟等领域展现了巨大潜力。它不仅能重现复杂的现实世界结构，还能拓展人类的想象力。然而，这些进展受限于3D数据集的稀缺性和高昂成本。尽管最近业界 [94, 117, 125] 开始创建大规模的专有3D资产，这些尝试需要大量的资金和运营资源。目前，学术界构建大规模3D数据集的成本依然过于

高昂。

## 动机

---

这种背景下，互联网视频提供了多视图图像的丰富来源。这些图像以不同的摄像机轨迹和传感器捕获，具有很高的扩展性、易获取性和低成本特点。因此，我们提出了以下问题：**模型能否通过大量的多视图图像学习到通用的3D先验？**

为了解决这一问题，本研究提出了一种无需姿态标注的视觉条件多视图扩散（MVD）模型，称为 **See3D**，实现开放世界的3D生成。

---

## 核心挑战

在利用互联网视频学习3D知识的过程中，我们需要解决以下核心问题：

1. **过滤3D感知相关的视频数据**，确保视频内容为静态场景，且摄像机视角有足够的变化。
  2. **从未明确标注几何信息和相机位姿的数据中学习通用的3D先验。**
- 

## 方法概述

我们的方法包括以下几个创新点：

- 提出了一种视频数据筛选管道，自动过滤动态内容或摄像机视角变化不足的视频。

- 构建了名为 **WebVi3D** 的数据集，包含1600万个视频片段，总计4.41年的视频内容。
- 设计了一种新的基于视觉条件的3D生成框架，支持在没有位姿标注的情况下进行高保真3D生成。

通过在单视图和稀疏视图重建基准上的比较实验，我们的方法展示了显著的零样本和开放世界生成能力。

---

## 关键贡献

1. **See3D**：一种可扩展的视觉条件多视图扩散模型，实现开放世界的3D创建，无需位姿标注。
2. **WebVi3D** 数据集：基于静态场景和多视图观察的自动化视频数据筛选管道。
3. 基于 **See3D** 的新型3D生成框架，支持复杂相机轨迹下的长序列生成。

## Related Work

### 从2D生成到3D生成

最近的3D生成方法主要受益于2D扩散模型的成功。这些方法可以分为以下两类：

#### 1. 优化3D表示：

通过最大化2D扩散先验的似然来优化3D表示，例如：

- 点云 (Point Clouds)
- 网格 (Meshes)
- 隐式场 (Implicit Neural Fields)

然而，2D先验很难直接转换为一致的3D表示，这导致了以下问题：

- **多视图不一致性**: 不同视角生成的结果无法保持一致。
- **全局几何误差**: 生成的整体几何结构可能存在偏差。

## 2. 基于变形-修补的管道:

- 使用离线深度估计器与基于2D扩散的修补模型相结合，逐步生成3D内容。
- 尽管这种方法适用于简单的3D场景，但在生成复杂场景时效果有限。

---

# 直接学习3D先验

为了更好地保留几何特性，近年来一些研究直接从3D数据中学习先验。这些方法可以分为以下几类：

## 1. 单视图/多视图输入:

- 使用编码-解码架构，直接生成3D表示（例如点云、体素网格等）。
- 这种方法通常无需额外的实例级优化。

## 2. 扩散模型预测3D表示:

- 利用扩散模型直接预测3D表示，如点云、网格和隐式场。
- 然而，大部分工作集中在物体级别的生成，对于场景级别的生成支持有限。

与这些方法不同，我们的工作通过大规模的互联网视频数据集训练模型，实现了物体级和场景级的3D生成。

---

## 学习多视图先验用于3D生成

多视图扩散模型（MVD）在最近的研究中得到了广泛关注，其主要优点在于：

- 生成能力：继承了2D扩散模型的生成能力。
- 多视图一致性：可以跨多个视角保持一致。

## 基于姿态的条件模型

传统的MVD模型依赖精确的相机位姿标注作为条件输入，例如：

- 相机外参 (Camera Extrinsics)
- 相对位姿 (Relative Poses)
- Plücker 光线 (Plücker Rays)

然而，这些模型的训练严重依赖昂贵的3D数据标注，限制了其扩展性。

## 基于视觉条件的模型

我们的工作引入了一种新颖的视觉条件方法，避免了对相机位姿的依赖：

- **无姿态条件**：通过视频中的2D视觉信号进行建模。
- **可扩展性**：支持大规模的无标注数据训练。

这种方法不仅减少了对昂贵标注数据的需求，还显著提高了模型在开放场景下的适应能力。

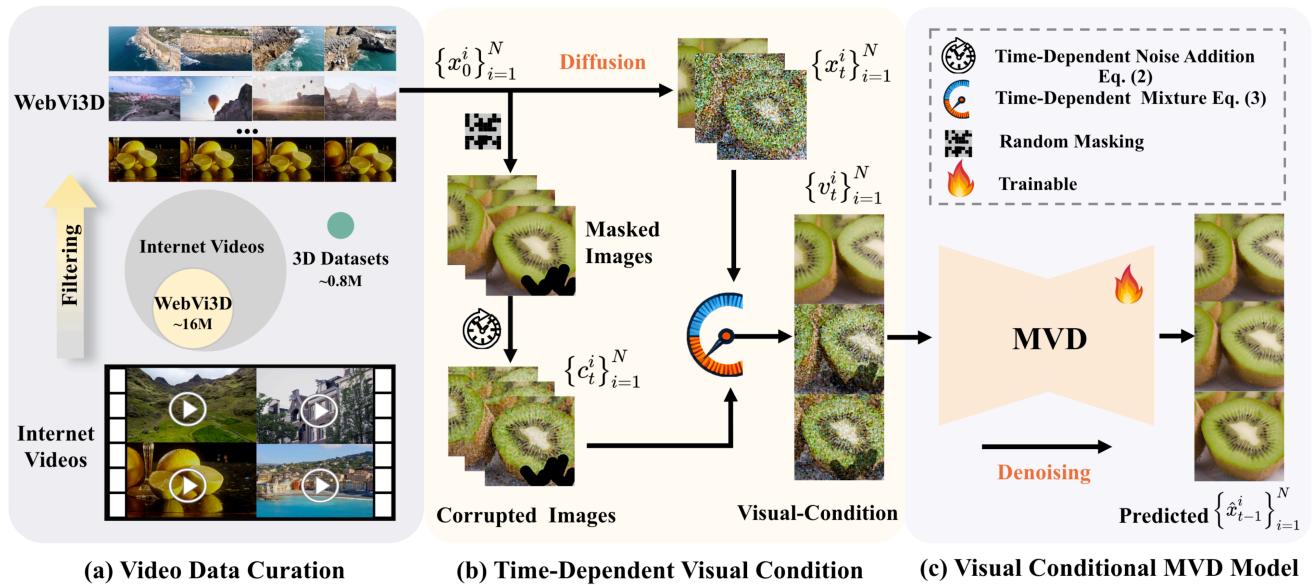


Figure 2. **Overview of See3D.** (a) We propose a four-step data curation pipeline to select multi-view images from Internet videos, forming the WebVi3D dataset, which includes  $\sim 16M$  video clips across diverse categories and concepts. (b) Given multiple views, we corrupt the original data into corrupted images  $c_t^i$  at timestep  $t$  by applying random masks and time-dependent noise. We then reweight the guidance of  $c_t^i$  and the noisy latent  $x_t^i$  for the diffusion model to form visual-condition  $v_t^i$  through a time-dependent mixture. (c) MVD model is capable of training at scale to generate multi-view images conditioned on  $v_t^i$ , without requiring pose annotations. Since  $v_t^i$  is a task-agnostic visual signal formed through time-dependent noise and mixture, it enables the trained model to robustly adapt to various downstream tasks.

# Method: Video Data Curation

## 数据筛选目标

高质量、规模化的视频数据是学习准确和鲁棒3D先验的基础。本部分的目标是从互联网上的大量视频中提取具有以下特性的3D感知视频：

1. **静态场景**: 确保在不同视角下几何结构一致。
2. **多视角变化**: 摄像机轨迹中存在足够的视角变化，以提供丰富的3D观察。

## 数据来源

我们从以下网站收集了 **25.48M** 个公开视频，总时长为 **44.98 年**：

- **Pexels**
- **Artgrid**
- **Airvuz**
- **Skypixel**

经过过滤后，得到 **2.30M** 视频，进一步分割为 **15.99M** 静态片段，构建了名为 **WebVi3D** 的数据集。

数据来源	原始视频数	筛选后视频数	筛选后片段数	筛选后总时长(小时)
Pexels	6.18M	0.61M	2.65M	9.96K
Artgrid	3.94M	0.54M	1.10M	8.77K
Airvuz	5.10M	0.54M	5.87M	8.72K
Skypixel	10.27M	0.61M	6.37M	8.82K

# 数据筛选流程

为了确保视频片段满足静态场景和多视角变化的要求，我们设计了以下四步筛选流程：

## 1. 时空下采样

对每段视频在时间和空间维度进行下采样以提高处理效率：

- 空间分辨率下采样至 **480p**。
- 时间下采样率设为 **2**。

## 2. 基于语义的动态识别

使用实例分割模型 **Mask R-CNN** 生成动态物体的运动掩码，筛选掉包含大量动态物体（如人类、动物、运动器材等）的帧。这一步确保筛选出的场景为静态场景。

## 3. 非刚性动态过滤

利用离线光流估计技术 (**Dense Optical Flow**) 获取非刚性运动区域的掩码，并分析这些掩码的位置以进一步判断视频是否包含动态内容。

## 4. 小视角变化过滤

利用像素跟踪技术 (**Pixel Tracking**) 计算关键点的运动轨迹，并估算最小外切圆的半径。对运动轨迹过小的视频进行过滤，确保视频中存在足够的视角变化。

---

## 筛选结果

最终，我们从 **25.48M** 个原始视频中提取了约 **320M** 张多视图图像。这些图像覆盖了静态场景和广泛的摄像机轨迹。

## 数据验证

随机选择 **10,000** 个视频片段进行人工标注，其中 **88.6%** 被标记为3D感知视频，证明了筛选流程的有效性。

---

## 未来扩展

随着互联网视频数据量的不断增长，该筛选管道能够持续获取更多的3D感知数据，使数据集的规模不断扩大。

# Method: Visual Conditional Multi-View Diffusion Model

---

## 前言

扩散模型通过前向扩散过程添加噪声，并通过学习逆向过程去除噪声，已经在生成任务中表现出强大的能力。本研究基于扩散模型，提出一种 **视觉条件多视图扩散模型 (MVD)**，无需相机姿态标注即可学习3D生成任务。

---

## 目标

目标是利用多视图预测生成指定相机轨迹下的新视图，确保与输入外观一致性。具体来说，我们的模型通过引入 **视觉条件 (Visual-Condition)**，从无标注的多视图视频中学习控制摄像机的移动，从而实现无姿态数据的高效训练。

目标优化函数为：

$$\mathbb{E}_{X_0, Y_0, \epsilon, t} [\|\epsilon_\theta(X_t, Y_0, V_t, t) - \epsilon\|_2^2] \quad (1)$$

其中：

- $X_t$  表示添加噪声后的潜变量。
  - $X_0 = \{x_i^0\}_{i=1}^N$  是从 WebVi3D 数据集中采样的一组多视图观测。
  - $Y_0 = \{y_i^0\}_{i=1}^S$  是随机选择的参考视图。
  - $G = \{g_i\}_{i=1}^L$  是目标生成图像。
  - $V_t$  是在时间步  $t$  上构造的视觉条件。
- 

## 可行性分析

为了避免昂贵的3D标注，所设计的视觉条件需满足以下要求：

1. 无需额外的3D标注。
  2. 任务无关性：适用于各种下游任务。
  3. 鲁棒性：能够处理不同任务之间的域差异。
- 

## 视觉条件的构造

### 1. 随机掩码 (Random Masking)

对目标图像  $G$  施加随机不规则掩码，减少模型对像素空间信号的直接依赖，同时保留参考图像  $Y_0$  的完整性以提供有效的外观信号。

---

### 2. 时间相关噪声 (Time-Dependent Noise)

在视频数据中添加时间相关噪声以近似高斯分布，具体定义如下：

$$C_t = \sqrt{\bar{\alpha}_{t'}}(1 - M)X_0 + \sqrt{1 - \bar{\alpha}_{t'}}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \quad (2)$$

其中：

- $\bar{\alpha}_{t'}$  为噪声调度器的方差。
- $M$  为掩码矩阵， $t' = f(t)$  是时间步  $t$  的非线性函数，防止噪声过大或过小。

### 3. 时间相关混合 (Time-Dependent Mixture)

引入时间步  $t$  的加权因子  $W_t$ , 对  $C_t$  和  $X_t$  进行混合:

$$V_t = [W_t \cdot C_t + (1 - W_t) \cdot X_t; M] \quad (3)$$

其中:

- $W_t \in [0, 1]$  是随时间步单调递减的权重因子。
  - $M$  是掩码矩阵, 确保参考图像  $Y_0$  的未掩码部分注入清晰的信息。
- 

## 模型架构

模型基于视频扩散模型设计, 移除了时间嵌入模块, 以通过视觉条件而非时间线索控制相机移动。具体改进包括:

- 对视频帧进行随机重排以打破时间顺序。
- 在每段视频中随机选择部分帧作为参考图像, 其余帧作为目标图像。
- 模型仅对目标图像计算损失, 优化目标如下:

$$\mathbb{E}_{X_0, Y_0, \epsilon, t} [\|\epsilon_\theta(X_t, Y_0, V_t, t) - \epsilon\|_2^2] \quad (4)$$

---

## 关键优势

1. **无姿态约束**: 模型仅依赖视觉条件进行训练, 无需昂贵的姿态标注。
2. **鲁棒性强**: 能够适应不同任务的域差异, 如多视图生成和基于掩码的3D编辑。

# Method: Visual Conditional Multi-View Diffusion Model

---

## 前言

---

扩散模型通过前向扩散过程添加噪声，并通过学习逆向过程去除噪声，已经在生成任务中表现出强大的能力。本研究基于扩散模型，提出一种 **视觉条件多视图扩散模型 (MVD)**，无需相机姿态标注即可学习3D生成任务。

---

## 目标

---

目标是利用多视图预测生成指定相机轨迹下的新视图，确保与输入外观一致性。具体来说，我们的模型通过引入 **视觉条件 (Visual-Condition)**，从无标注的多视图视频中学习控制摄像机的移动，从而实现无姿态数据的高效训练。

目标优化函数为：

$$\mathbb{E}_{X_0, Y_0, \epsilon, t} [\|\epsilon_\theta(X_t, Y_0, V_t, t) - \epsilon\|_2^2] \quad (5)$$

其中：

- $X_t$  表示添加噪声后的潜变量。
- $X_0 = \{x_i^0\}_{i=1}^N$  是从 WebVi3D 数据集中采样的一组多视图观测。
- $Y_0 = \{y_i^0\}_{i=1}^S$  是随机选择的参考视图。

- $G = \{g_i\}_{i=1}^L$  是目标生成图像。
  - $V_t$  是在时间步  $t$  上构造的视觉条件。
- 

## 可行性分析

为了避免昂贵的3D标注，所设计的视觉条件需满足以下要求：

1. 无需额外的3D标注。
  2. 任务无关性：适用于各种下游任务。
  3. 鲁棒性：能够处理不同任务之间的域差异。
- 

## 视觉条件的构造

### 1. 随机掩码 (Random Masking)

对目标图像  $G$  施加随机不规则掩码，减少模型对像素空间信号的直接依赖，同时保留参考图像  $Y_0$  的完整性以提供有效的外观信号。

---

### 2. 时间相关噪声 (Time-Dependent Noise)

在视频数据中添加时间相关噪声以近似高斯分布，具体定义如下：

$$C_t = \sqrt{\bar{\alpha}_{t'}}(1 - M)X_0 + \sqrt{1 - \bar{\alpha}_{t'}}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \quad (6)$$

其中：

- $\bar{\alpha}_{t'}$  为噪声调度器的方差。
- $M$  为掩码矩阵,  $t' = f(t)$  是时间步  $t$  的非线性函数, 防止噪声过大或过小。

### 3. 时间相关混合 (Time-Dependent Mixture)

引入时间步  $t$  的加权因子  $W_t$ , 对  $C_t$  和  $X_t$  进行混合:

$$V_t = [W_t \cdot C_t + (1 - W_t) \cdot X_t; M] \quad (7)$$

其中:

- $W_t \in [0, 1]$  是随时间步单调递减的权重因子。
- $M$  是掩码矩阵, 确保参考图像  $Y_0$  的未掩码部分注入清晰的信息。

## 模型架构

模型基于视频扩散模型设计, 移除了时间嵌入模块, 以通过视觉条件而非时间线索控制相机移动。具体改进包括:

- 对视频帧进行随机重排以打破时间顺序。
- 在每段视频中随机选择部分帧作为参考图像, 其余帧作为目标图像。
- 模型仅对目标图像计算损失, 优化目标如下:

$$\mathbb{E}_{X_0, Y_0, \epsilon, t} [\|\epsilon_\theta(X_t, Y_0, V_t, t) - \epsilon\|_2^2] \quad (8)$$

# 关键优势

---

1. 无姿态约束：模型仅依赖视觉条件进行训练，无需昂贵的姿态标注。
2. 鲁棒性强：能够适应不同任务的域差异，如多视图生成和基于掩码的3D编辑。

## Time-dependent Visual Condition

---

### 背景

为了从视频中学习摄像机的运动控制，我们设计了一种时间相关的视觉条件(**Time-dependent Visual Condition**)，通过结合随机掩码、时间相关噪声和混合策略，增强模型的鲁棒性和泛化能力。

---

### 方法

#### 1. 随机掩码 (Random Masking)

目标图像  $G$  被随机掩盖部分像素，以减少显性视觉信号的直接依赖，同时参考图像  $Y_0$  保持完整。这样可以帮助模型：

- 专注于捕获几何和上下文信息。
- 在目标图像不可见的区域生成一致的内容。

公式表示：

$$M = \{m_0 : S \cup m_{S+1:N}\} \quad (9)$$

其中：

- $m_0 : S$  是参考图像  $Y_0$  的零矩阵（未掩码）。
  - $m_{S+1:N}$  是对目标图像  $G$  应用的随机掩码。
- 

## 2. 时间相关噪声 (Time-dependent Noise)

通过向掩码后的图像添加噪声，将其分布调整为高斯分布。公式如下：

$$C_t = \sqrt{\bar{\alpha}_{t'}}(1 - M)X_0 + \sqrt{1 - \bar{\alpha}_{t'}}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \quad (10)$$

其中：

- $X_0$  是原始输入。
- $\bar{\alpha}_{t'}$  是噪声调度器的方差。
- $t' = f(t)$  是时间步  $t$  的非线性映射函数，确保  $t' < t$ ，以防止噪声过度削弱视觉信号。

### 设计非线性函数 $f(t)$

$f(t)$  的作用是调节噪声强度，防止在较大时间步下信号完全丢失。常见设计包括：

- 单调递增函数（例如线性或指数）。
-

### 3. 时间相关混合 (Time-dependent Mixture)

为了平衡噪声信号和潜变量  $X_t$  的贡献，引入时间相关的加权混合策略：

$$V_t = [W_t \cdot C_t + (1 - W_t) \cdot X_t; M] \quad (11)$$

其中：

- $W_t \in [0, 1]$  是随时间步  $t$  单调递减的权重因子。
- $C_t$  是掩码后的噪声信号。
- $X_t$  是噪声潜变量。

混合策略的设计目的在于：

1. 在早期时间步更多依赖像素空间信号  $C_t$ ，以提供初始外观信息。
2. 在后期时间步逐步转向依赖潜变量  $X_t$ ，以捕获复杂的上下文信息。

---

## 最终视觉条件表达式

---

综合掩码和噪声，最终的时间相关视觉条件定义为：

$$V_t = [W_t \cdot C_t + (1 - W_t) \cdot X_t; M] \quad (12)$$

- 参考图像部分直接注入清晰信息。
- 目标图像部分结合了噪声和潜变量的信息，增强模型对不可见区域的生成能力。

# 优势

---

1. **动态适应性**: 通过时间相关的混合策略, 动态平衡不同来源的信号。
  2. **域鲁棒性**: 在处理不同任务时, 模型能够适应域间差异, 例如任务特定的像素提示和视频帧之间的分布差异。
  3. **无标注依赖**: 仅需像素空间的提示, 无需依赖昂贵的3D姿态标注。
- 

## 示例

---

在具体实现中,  $W_t$  和  $t'$  的选择直接影响视觉条件的效果。例如:

- $W_t = e^{-kt}$  (指数衰减)。
- $f(t) = \sqrt{t}$  (平方根映射)。

这些选择可以根据实际任务进行调整, 以平衡噪声与潜变量的权重。

# Model Architecture

---

## 总体架构

---

模型的核心基于视频扩散模型, 设计目标是通过**视觉条件 (Visual-Condition)** 控制摄像机运动, 而无需依赖时间嵌入或3D姿态标注。模型架构的关键组件包括:

1. **去除时间嵌入**: 使模型通过视觉条件而非时间线索控制运动。
2. **随机帧重排**: 打破时间顺序, 模拟多视图无序输入。

3. 参考视图与目标视图划分：利用参考视图提供外观信号，目标视图进行优化。
- 

## 数据预处理

- 输入数据：每段视频中随机选择  $N$  帧作为多视图观测

$$X_0 = \{x_i^0\}_{i=1}^N.$$

- 参考视图与目标视图划分：

- $Y_0 = \{y_i^0\}_{i=1}^S$  是参考视图，提供完整外观信号。

- $G = \{g_i\}_{i=1}^L$  是目标视图，用于生成新视角。

优化目标为：

$$\mathbb{E}_{X_0, Y_0, \epsilon, t} [\|\epsilon_\theta(X_t, Y_0, V_t, t) - \epsilon\|_2^2] \quad (13)$$

---

## 模型模块

### 1. 扩散模型核心模块

模型基于扩散过程，通过逐步去噪生成目标图像。

- 前向扩散过程：

$$X_t = \sqrt{\bar{\alpha}_t} X_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

- $X_t$  是时间步  $t$  添加噪声后的潜变量。

- $\bar{\alpha}_t$  是噪声调度器的方差。

- 逆向扩散过程：

学习去噪函数  $\epsilon_\theta$ , 优化目标为:

$$\mathbb{E}_{X_0, \epsilon, t} [\|\epsilon_\theta(X_t, Y_0, V_t, t) - \epsilon\|_2^2]$$

---

## 2. 自注意力层 (Self-Attention Layers)

自注意力机制用于捕获输入多视图帧之间的全局上下文关系。

- 多头注意力机制 (Multi-Head Attention):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

- $Q, K, V$  分别为查询、键和值矩阵。
- $d_k$  是键向量的维度。

自注意力模块增强了多视图输入的全局一致性。

---

## 3. 跨注意力层 (Cross-Attention Layers)

跨注意力层通过视觉条件  $V_t$  调控目标图像的生成过程。

- 查询来自目标帧:  $Q_t = X_t$ 。
- 键和值来自视觉条件:  $K_t, V_t = \text{Visual-Condition}$ 。

生成的新视图通过整合视觉条件中的多视图信息保持一致性。

---

## 4. 时间嵌入的移除

为了使模型通过视觉条件而非时间趋势控制摄像机运动：

- 去除了时间步  $t$  的直接嵌入。
  - 通过视觉条件  $V_t$  显式控制生成过程。
- 

## 5. 损失函数

损失函数仅对目标视图计算，优化模型在去噪过程中的性能：

$$\mathbb{E}_{X_0, Y_0, \epsilon, t} [\|\epsilon_\theta(X_t, Y_0, V_t, t) - \epsilon\|_2^2] \quad (14)$$

---

## 优化流程

1. 随机选择  $N$  帧作为输入数据  $X_0$ 。
  2. 划分参考视图  $Y_0$  和目标视图  $G$ 。
  3. 构造时间相关视觉条件  $V_t$ 。
  4. 扩散模型通过视觉条件  $V_t$  控制目标图像  $G$  的生成。
  5. 仅对目标图像  $G$  计算损失，优化去噪函数  $\epsilon_\theta$ 。
- 

## 关键改进

1. **无姿态标注**：通过视觉条件取代相机位姿，解决了标注昂贵的问题。
2. **强一致性**：自注意力和跨注意力模块增强了多视图生成的全局一致性。

3. 灵活适配：架构可扩展至多种任务，包括单视图生成和稀疏视图重建。

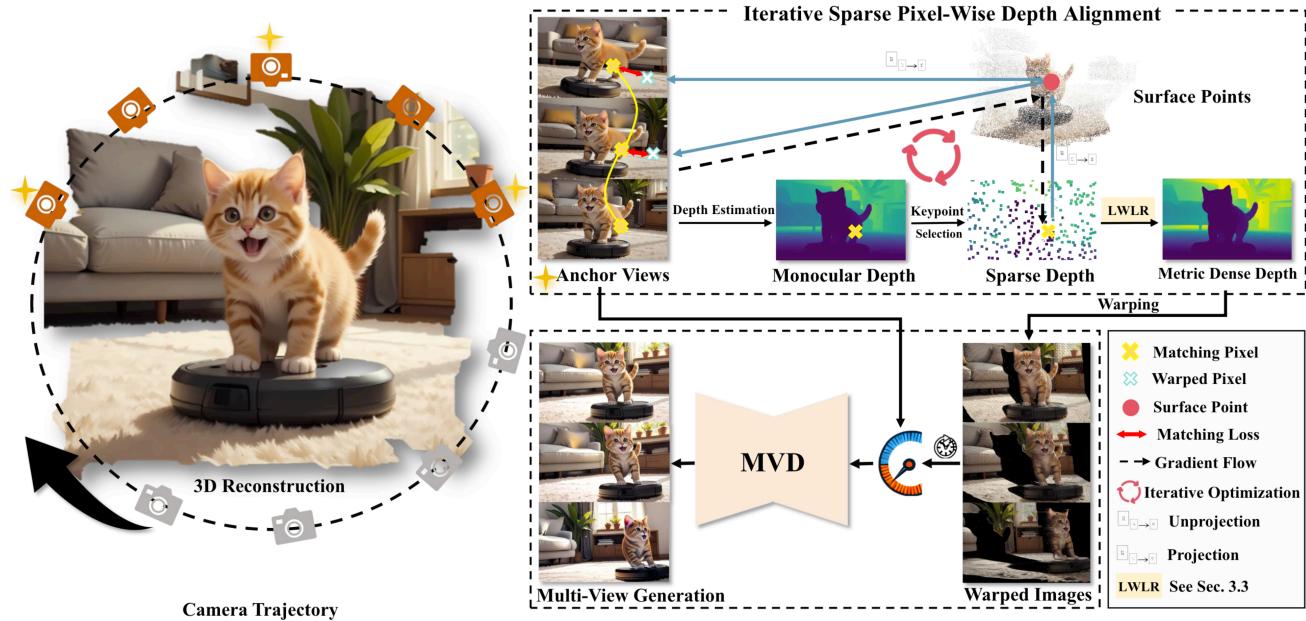


Figure 4. **See3D for Multi-View Generation**: From iteratively generated views (brown camera), we randomly select a few anchor views (yellow stars) to guide the generation of target views along the gray camera trajectory. Keypoint matching is first performed to establish correspondences between the anchor views. Next, monocular depth estimation is applied to the latest anchor view, followed by our *Iterative Sparse Pixel-Wise Depth Alignment* to refine the depth and recover a dense map. This dense depth is then used to warp images along the gray camera viewpoints. Subsequently, the warped images and anchor images are combined and processed according to Eq.2 and Eq.3, without random masking, forming the *visual-condition*, which guides MVD model to produce 3D-consistent target views. Finally, the gray camera turns to brown, guiding multi-view generation in the next iteration.

# Visual Conditional 3D Generation

## 概述

本部分提出了一种基于视觉条件的 3D生成框架，支持复杂摄像机轨迹下的长序列新视角生成。该框架以 **See3D** 为核心，结合多视图生成、深度对齐和迭代优化策略，实现高保真度的3D场景重建。

# 生成流程

---

该框架的主要步骤包括：

1. 深度估计与对齐：对已有视图的深度进行校准，以提高几何一致性。
  2. 图像变形与视觉条件构造：通过变形操作生成新视角的视觉提示。
  3. **See3D生成**：利用视觉条件生成与多视图一致的新视角图像。
  4. 迭代优化：通过多轮生成进一步丰富视角信息。
- 

## 1. 深度估计与对齐

### 1.1 像素级深度标定

通过关键点匹配实现高精度的像素级深度校准：

- 使用关键点检测和匹配（例如 **SuperPoint** 和 **LightGlue**），找到跨视图的对应点对  $\{m_n, m_i\}_k$ 。
- 对每对关键点  $k$ ，优化尺度  $\alpha_k$  和偏移  $\beta_k$  以校准深度：

$$\alpha_k^*, \beta_k^* = \arg \min_{\alpha_k, \beta_k} \|d_n^k K_i T_i T_n^{-1} K_n^{-1} m_n^t - m_i^t\|_2^2 \quad (15)$$

- $d_n^k$  是源视图中关键点的深度值。
- $K$  和  $T$  分别为相机的内参和外参矩阵。

### 1.2 全局深度恢复

利用局部加权线性回归 (**LWLR**) 从稀疏深度恢复密集深度图：

$$\min_{\beta_{u,v}} (d_n^* - X\beta_{u,v})^T W_{u,v} (d_n^* - X\beta_{u,v}) + \lambda S_{\text{shift}}^2 \quad (16)$$

其中：

- $W_{u,v}$  是高斯加权矩阵，定义为：

$$w_i = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\text{dist}_i^2}{2b^2}\right) \quad (17)$$

- $\text{dist}_i$  是指导点与目标点之间的欧几里得距离。
  - $\lambda$  是正则化参数，用于平滑解。
- 

## 2. 图像变形与视觉条件构造

### 2.1 图像变形

通过变形操作生成目标视角的初步视觉提示：

$$I_j = \Pi_{n \rightarrow j}(D_n) \quad (18)$$

- $\Pi_{n \rightarrow j}$  是像素从源视图到目标视图的投影操作，定义为：

$$\Pi_{n \rightarrow j}(d_n) = d_n K_j T_j T_n^{-1} K_n^{-1} \quad (19)$$

- 变形后图像可能出现空洞区域，用二值掩码  $M_j$  表示未填充区域。

## 2.2 视觉条件构造

结合变形后的图像和原始参考视图，构造时间相关视觉条件：

$$V_t = [W_t \cdot I_j + (1 - W_t) \cdot X_t; M_j] \quad (20)$$

- $W_t$  是时间步  $t$  的权重因子。
  - $M_j$  是掩码矩阵。
- 

## 3. See3D生成

通过 **See3D** 生成与多视图一致的新视角图像：

$$I_j = \text{See3D}(I_j, M_j, \{I_0, I_k\}) \quad (21)$$

- $I_k$  是随机选择的锚点视图，用于指导多视图生成的一致性。
- 

## 4. 迭代优化

逐步迭代生成多视角图像，更新生成序列：

1. 使用最新生成的视图作为锚点视图。
  2. 根据预定义相机轨迹  $\{T_i\}$ ，重复深度校准、变形和生成步骤。
  3. 持续优化生成视图的全局几何一致性。
- 

## 3D重建

使用 **3D Gaussian Splatting (3DGS)** 进行场景重建：

1. **训练目标**: 最小化光度损失、结构相似性损失 (SSIM) 和感知损失 (LPIPS):

$$\mathcal{L} = \mathcal{L}_{\text{photo}} + \mathcal{L}_{\text{SSIM}} + \mathcal{L}_{\text{LPIPS}} \quad (22)$$

2. **相机优化**: 对生成图像和相机位姿进行联合优化，确保视角一致性。

## 优势总结

1. **多视图一致性**: 通过深度对齐和视觉条件增强了生成图像的几何一致性。
2. **任务泛化性**: 框架能够适应不同的生成任务，包括单视图生成、稀疏重建和场景编辑。
3. **高效扩展性**: 通过迭代优化和视觉条件支持长序列视角生成。

## Conclusion

## 总结

本研究提出了一个可扩展的3D生成框架，旨在通过数据规模化和视觉条件的创新设计，突破传统3D生成方法对昂贵3D数据标注的依赖，推动3D生成技术在开放世界场景中的应用。研究的主要贡献包括：

1. **WebVi3D 数据集**:

- 构建了一个大规模、自动化筛选的视频数据集，包含 **15.99M** 静态视频片段，总时长 **4.41 年**，涵盖多种场景与摄像机轨迹。
- 数据集通过自动化管道筛选，能够从互联网视频中持续扩展高质量 3D 感知数据。

## 2. See3D 模型：

- 提出了基于视觉条件的多视图扩散模型 (**MVD**)，实现了无姿态约束下的大规模训练。
- 视觉条件通过随机掩码和时间相关噪声的设计，有效引导模型学习多视图一致性和摄像机运动控制。

## 3. 视觉条件3D生成框架：

- 基于视觉条件构建了一个新颖的3D生成框架，支持复杂摄像机轨迹下的长序列生成。
- 框架结合深度校准和迭代优化，确保生成的3D场景具有高几何保真度和视图一致性。

---

# 实验结果

---

实验结果表明：

- 在单视图生成和稀疏视图重建任务中，**See3D** 展现出卓越的零样本和开放世界生成能力。
- 与基准模型相比，See3D 显著提升了生成质量（PSNR、SSIM、LPIPS 等指标）。

# 启示与未来工作

---

## 1. 3D生成的可扩展性：

- 本研究表明，通过视觉条件和大规模视频数据，模型可以有效绕过传统3D生成方法的标注瓶颈。
- 数据规模的进一步扩展将提升3D生成的上限。

## 2. 对3D社区的贡献：

- WebVi3D 和 See3D 为 3D 研究提供了一种低成本、高效的解决方案。
- 通过开放的技术框架，学术界能够追求与闭源3D生成解决方案的性能接近甚至超越。

## 3. 未来方向：

- 优化视觉条件的构造方法，提升模型对更复杂场景的适应能力。
  - 探索跨模态生成任务，例如结合文本或其他输入模态进行3D生成。
  - 利用 WebVi3D 数据集的动态扩展能力，持续推动开放世界 3D 生成的性能提升。
- 

# 结语

本研究从数据规模化和框架创新两个角度，为开放世界 3D 生成提供了一条全新的技术路径。通过降低 3D 数据采集和标注的门槛，我们期望本研究能够推动 3D 技术在虚拟现实、娱乐和仿真等领域的广泛应用，并激发更多针对大规模数据的 3D 生成研究。