下一视角预测驱动的一致性图生三维物体方法

张旭迎 [1*] 周宇鹏 [1*] 汪楷 [1] 王一凯 [2] 李震 [1]
焦少慧 [3] 周大权 [3] 侯淇彬 [1†] 程明明 [1]

[1] 南开大学 [2] 清华大学 [3] 字节跳动

项目主页:https://github.com/HVision-NKU/AR123

图 1. 我们的 AR-1-to-3 模型生成的高质量三维资产示例展示。

Abstract

新视角合成（*Novel View Synthesis, NVS*）是图像生成三维模型任务的基石。然而，现有方法仍然难以保持生成视角与输入视角之间的一致性，尤其是存在显著的相机姿态差异时，经常生成质量较差的三维几何和纹理。我们将这一问题归因于现有方法对所有目标视角赋予了相同的优先级；根据我们的实证观察，距离输入视角较近的目标视角通常具有更高的保真度。基于此启发，我们提出了一种基于扩散模型的新颖的下一视角预测范式——*AR-1-to-3*，该方法首先生成与输入视角接近的图像，并将其作为上下文信息，用于逐步合成更远距离的视角。为了将已生成的视角连续性编码为下一视角预测所需的局部与全局条件，我们相应地设计了一个堆叠式局部特征编码策略（*Stacked-LE*）和一个基于 *LSTM* 的全局特征编码策略（*LSTM-GE*）。大量实验表明，我们的方法显著提升了生成视角与输入视角之间的一致性，从而生成高保真度的三维资产。代码与预训练权重将公开发布。

1. 绪论

从单视图图像合成三维物体是计算机视觉与图形学领域一项基础且持久的研究课题。这项任务具有挑战性，它不仅需要重建三维物体的可见区域，还需对不可见部分进行合理的外推。为了解决这些挑战，研究人员尝试使用已经图像生成任务中取得显著成果的扩散模型 [19, 29, 33, 61]，借助其强大的先验能力合成新视角图像 [4, 22, 25, 31]。

现有方法大致可以分为两类范式。具体来说，以

* 共同一作。
† 通讯作者。

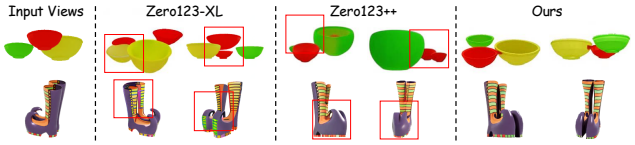

图 2. 在相机姿态差异较大的情况下，不同范式生成的新视角示例对比。

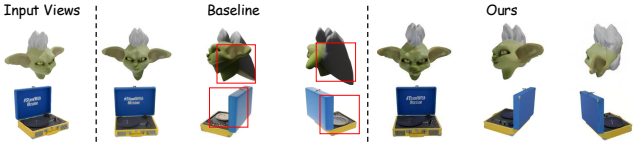

图 3. Zero123++ 基线方法与我们提出的 AR-1-to-3 在近距与远距相机视角下生成的新视角示例对比。

Zero123 [23] 为代表的方法，通过在 <*input_view, camera_pose, output_view*> 三元组的三维渲染图上微调预训练的扩散模型，依据单视图图像和一组相机姿态生成三维物体的多个离散视角图像。另一类方法如 Zero123++ [35] 及其后续工作 [17, 21, 40, 54]，则认为不同相机姿态指定的新视角之间是独立的。这类方法将所有目标视角和其对应的相机姿态合并为网格图像，以实现联合分布建模，并通过扩散模型同时生成多个视角。尽管这些方法在许多场景中取得成功，但在新视角与输入图像保持一致性方面仍存在问题，尤其当相机姿态差异较大时，常出现不一致现象，如图 2中的碗与鞋的示例所示。

造成该问题的一个根本原因是，这些方法对所有相机姿态对应的新视角赋予了相同的生成优先级。我们认为，应当优先生成与输入视角相近的目标视角，因为它们通常具有更高的生成保真度；而姿态变化较大的视角则更具挑战性，如图 3所示。为此，我们重新思考了新视角生成的方式，提出了一种新的范式 AR-1-to-3，以自回归方式逐步生成所有目标视角，优先生成距离较近的视角，并将其作为上下文信息用于后续远距离视角的生成。

我们的方法采用了 Zero123++ [35] 提出的 3×2 网格图像生成策略。需要指出的是，这六个目标视角之间存在潜在的序列关系，即网格图像中相邻的行具有相同的俯仰角，且方位角之间固定间隔为 120°。这种序列特性使得我们的 AR-1-to-3 可以从第一行视角的生成开始，并以自回归方式逐步生成剩余的视角。在每一步迭代中，来自不同俯仰角的目标视角可以进行信息交换，前一步生成的图像也会被用作当前视角生成的参考。

为了对已生成的部分视角序列进行编码并为下一步提供参考，我们分别设计了两种图像条件编码策略，即用于局部调控的 Stacked-LE 和用于全局调控的 LSTM-GE。在 Stacked-LE 策略中，去噪 UNet 模型将先前生成的视角编码为堆叠式嵌入，作为像素级空间引导，用于调整自注意力模块中的 key 和 value 矩阵，从而完成当前目标视角的去噪。在 LSTM-GE 中，视角序列根据俯仰角被分为两个子组，其特征向量分别通过两个 LSTM 模块编码，作为当前视角的高层语义条件。

我们在三维基准数据集 Objaverse [7] 和两个域外数据集 Google Scanned Objects [8] 与 OmniObject3D [53] 上评估了 AR-1-to-3 的性能。通过在多视角扩散模型中引入自回归生成方式，并结合所提出的 Stacked-LE 与 LSTM-GE 策略，我们的方法能够实现一致且准确的新视角合成，进而生成高质量的三维资产，如图 1所示。实验结果还表明，与现有先进的新视角合成方法和图像到三维生成方法相比，AR-1-to-3 具有明显优势。

我们的主要贡献如下：

- 我们提出了一种用于三维物体生成的自回归式下一视角预测框架 AR-1-to-3，可以实现从近到远逐步生成目标视角；
- 我们设计了 Stacked-LE 和 LSTM-GE 两种特征编码策略，用于对已生成的视角序列进行编码，并为扩散模型提供局部与全局调控条件；
- 在多个大规模三维数据集上的定量与定性实验表明，我们的方法能够生成比以往方法更加一致的多视角图像，并得到高质量的三维资产。

2. 相关工作

2.1. 用于三维生成的二维扩散模型

在大规模二维图像数据集上预训练的扩散模型 [11, 16, 29, 33] 不仅在高质量图像生成方面表现出色，且具有强大的零样本泛化能力。近年来，研究人员不断努力将二维扩散模型的这些优势迁移到三维生成任务中。

DreamFusion [30]、SJC [49] 和 Fantasia3D [2] 提出将渲染视图输入预训练的二维扩散模型，并对每个输入视图进行单独优化，以此蒸馏其知识，但经常出

现如颜色过饱和与“多面”等伪影问题。Zero123 [23] 首次提出了开放世界的单图像生成三维的框架，通过微调扩散模型，在输入视角和一组离散相机姿态条件下合成新视角图像。ImageDream [50] 采用与 MVDream [37] 相同的世界坐标系来恢复三维几何形状。Magic123 [31] 将 Zero123 的三维先验与稳定扩散模型的二维先验相结合，以提升生成三维网格的质量。One-2-3-45 [22] 使用 Zero123 生成多视角图像，并将其提升到三维空间以辅助三维网格生成。Consistent123 [20] 和 MVD-Fusion [14] 等方法在 Zero123 的基础上引入了边界、深度等附加先验以提升表现。此外，越来越多的研究开始关注多视角生成结果之间的一致性问题。SyncDreamer [24] 引入三维感知的注意力机制，以关联不同视角间的对应特征。MVDiffusion [41] 则通过权重共享的多分支 UNet 和考虑对应关系的注意力机制，实现并行生成多视角图像。最近，Cycle3D [42] 和 Free3D [60] 通过同时生成多个视角建模其联合分布。Zero123++ [35] 进一步提出将围绕三维物体的六个目标视角排列为一个网格图像的策略。该策略被后续多个方法广泛采用，如 One-2-3-45++ [21]、Instant3D [17] 和 InstantMesh [54] 等。

与上述高度依赖二维扩散先验的方法不同，我们提出的 AR-1-to-3 受到人类思维方式的启发，更加注重当前物体的上下文信息。与我们方法最接近的是 Cascade-Zero123 [4]。该方法首先使用多视角扩散模型生成大量额外视角，然后与输入图像一起输入另一个扩散模型，以生成指定目标视角。与其不同的是，我们的方法 AR-1-to-3 明确考虑了目标视角与输入图像之间的关系，并利用扩散模型建模它们之间的潜在序列结构。

2.2. 自回归生成

自回归机制通常用于时间序列分析与预测，其基本思想是序列中当前的数值可以由其前序值决定。基于这一思想，研究者发展出了一系列经典的序列建模方法 [12, 28, 36, 47]。近年来，越来越多的研究尝试将这一自回归模式扩展到多个领域。

PixelRNN [45] 是最早通过建模像素间依赖关系以生成细节丰富的高质量图像的方法之一。VQ-VAE [46] 通过引入码本机制改进了离散表征的学习方式，实现了图像的高效编码与解码。VQ-GAN [9] 则结合 Transformer 架构以序列方式建模图像视觉块，并在训练中引入对抗损失。Parti [55] 提出路径式自回归模型，将图像生成建模为序列到序列的任务，生成高保真的写实图像。VAR [43]、LlamaGen [39] 和 Infinity [10] 借助多模态大模型提升图像生成规模。此外，还有部分工作 [1, 15, 34, 48, 62] 将扩散模型与序列建模策略结合，以实现时间一致性更强的视频生成。近期，MeshGPT [38]、MeshAnything [5] 和 MeshXL [3] 等方法提出以自回归方式逐面生成艺术家创建的三维网格。TAR3D [58] 和 SAR3D [6] 对三维潜变量进行量化，并通过下一个 token 预测策略生成三维物体。

在本工作中，我们观察到 Zero123++ 的目标视角可以根据相机姿态的等间隔划分为若干生成步骤，呈现出潜在的序列性。因此，我们提出以自回归方式逐步生成多个新视角图像。

3. 本文方法

3.1. 预备知识

为了更好地理解 AR-1-to-3 的结构设计，我们首先介绍本文所采用的基础多视角扩散模型 Zero123++ [35] 。

多视角生成. 为了建模多个新视角之间的联合分布，Zero123++ 提出将六个目标视角以 3×2 的布局拼接为一张网格图像。需要注意的是，这些目标视角是通过一组固定的相对方位角与绝对俯仰角获得的。具体而言，它们由交替的俯仰角（20° 向下与 −10° 向上）与从输入视角起始、每次递增 60° 的方位角组合而成。

稳定扩散模型. Zero123++ 选用了稳定扩散模型作为生成模型，因为其为开源模型，并在大规模互联网图像数据集上完成了预训练。该模型从自然图像中学习到的几何先验被用于在图像与相机条件下合成新视角图像。稳定扩散模型在预训练自动编码器的潜空间中执行扩散过程，自动编码器的编码器与解码器分别表示为 $\mathcal{E}(\cdot)$ 和 $\mathcal{D}(\cdot)$。在扩散的第 t 步，去噪 UNet $\epsilon_\theta(\cdot)$ 的微调目标可定义为：

$$\mathcal{L} = \mathbb{E}_{z\sim\mathcal{E}(x),y,t,\epsilon\sim\mathcal{N}(0,I)} ||\epsilon - \epsilon_\theta(z_t, t, c_\theta(y))||_2^2, \quad (1)$$

其中 x 表示目标网格图像，在扩散过程中被加入高斯噪声 ϵ 并扰动为 z_t，而 $c_\theta(y)$ 表示由条件图像 y 编码而来的嵌入特征。

图像条件. Zero123++ 中的图像调控策略 $c_\theta(y)$ 可分为局部调控与全局调控两类。局部调控主要用于建模

输入视图与目标视图之间的像素级空间对应关系。具体做法是采用一种参考注意力操作的变体 [56]，首先在输入图像上运行去噪 UNet，并将其中自注意力模块的 key 与 value 矩阵附加至目标视角去噪过程的对应层。注意：为了使 UNet 聚焦于当前噪声层级的有效特征，参考图像中也加入了与目标图像相同层级的高斯噪声。

全局调控方面，Zero123++ 首先将输入图像的 CLIP [32] 图像嵌入乘以一组可学习的全局权重，再与空文本的 CLIP 文本嵌入相加，最终作为去噪 UNet 中交叉注意力的高层语义引导。

3.2. AR-1-to-3

现有方法如 Zero123 [23]、One-2-3-45 [22] 通常从单视图图像与一组相机姿态出发，生成多个离散视角图像；而 Zero123++ [35]、ImageDream [50] 等则采用网格布局，同时生成多个相机条件下的视角图像。尽管这些方法在多个场景中表现良好，但仍易出现目标视角与输入图像在几何与纹理上的不一致问题。我们认为，其对所有目标视角赋予相同优先级，且在生成过程中未充分利用当前物体的上下文信息，是导致该问题的关键原因。

本文的核心思想是以“下一视角预测”的方式逐步生成目标视图，使得较早生成的近距离视图可以为后续远距离视图的生成提供补充信息。图 4 展示了 AR-1-to-3 的端到端结构。我们遵循 Zero123++ 的范式，生成 6 个特定相机条件下的目标视角，但不同的是，我们并非一次性同时生成所有视角，而是分阶段逐步生成。此外，Zero123++ 已经证明同时生成多个目标视角有助于准确建模其联合分布。因此，我们的方法将每一步生成对应于 3×2 网格布局中的一行，其中包含两个目标视角，俯仰角不同，方位角相差 60°。每一步之间的相机姿态间隔为固定的 120° 方位角，这种结构天然适合用于下一视角预测。由此，每一步中的目标视角可以相互交换信息，且前一步生成的视角可作为附加条件，用于指导当前步骤的目标视角生成。

我们通过设计两种图像条件策略，实现了这种“下一视角预测”，这两种策略用于编码视角序列信息，以微调扩散模型的生成过程。这两种策略分别称为堆叠式局部特征编码（Stacked-LE）与长短期全局特征编码（LSTM-GE），对应于 Zero123++ 中的局部与全局图像调控机制。它们的优化目标仍由公式 (1) 表示，而具体的图像调控策略 $c_\theta(y)$ 将在后续的 第 3.3节与 第 3.4节中详细介绍。

通过多阶段自回归生成，我们的 AR-1-to-3 最终得到 6 个目标视角图像，这些图像将被输入至一个稀疏视角的大模型用于重建三维物体。本文中，我们采用预训练的 InstantMesh [54] 作为三维重建模型，其通过基于 Transformer 的架构 [13] 将多视角图像编码为 triplane 特征，并使用多层感知机预测体渲染所需的点颜色与密度。

3.3. 堆叠式局部特征编码

本节介绍我们的方法如何将输入图像与已生成的部分目标视角序列的潜在特征编码为局部条件，以用于参考注意力操作，从而生成当前步骤的目标视角。需要注意的是，去噪 UNet 模型具有多层结构，且不同自注意力层的隐藏维度可能不一致。这使得难以通过一个统一的网络，在所有位置上对条件视角序列的潜在特征进行编码。考虑到这些参考特征与自注意力层中的注意力表示来自 UNet 的相同位置，具有一致的空间和通道维度，我们自然地采用在空间维度上进行堆叠的方式，将其统一编码为一个表示。这种策略有两个明显优势：1）它可以在自注意力层中编码任意数量的参考特征；2）它允许参考特征直接作为注意力模块的输入，从而无需额外设计即可复用原有的网络权重参数。我们将此局部特征编码策略称为 Stacked-LE，其结构如 图 4 左下角所示。

形式上，在第 k 个自回归步骤中，已生成了 $2k-1$ 个参考视角，我们目标是预测第 $2k$ 和第 $2k+1$ 个目标视角，其中 k 的取值范围为 1 到 3。按照 Zero123++ [35] 的做法，我们首先将每个参考视角分别输入去噪 UNet，并记录其在自注意力模块中的 key 和 value 矩阵。随后，我们再进行一次前向传播，对当前步骤的目标视角进行去噪。在该过程中，将每一层记录下的参考特征在空间维度上进行堆叠，并用于修改该层自注意力模块中的 key 和 value 矩阵，公式如下：

$$s_i^* = \text{Concat}([e_i^1, e_i^2, ..., e_i^{2k-1}, s_i]), \tag{2}$$

其中 $s_i \in \mathbb{R}^{B \times L \times D_i}$ 表示第 i 层自注意力的 key 或 value 矩阵，e_i^j 表示第 j 个参考视角在该层记录的嵌入特征。变量 B、L 与 D_i 分别表示 batch 大小、token

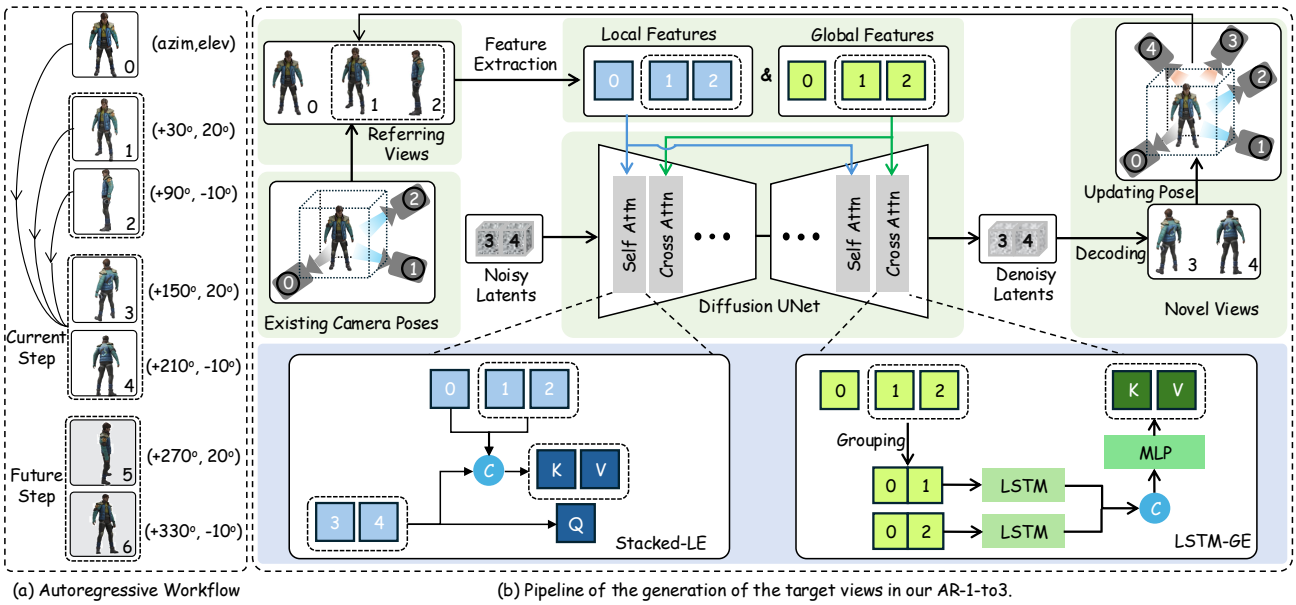

图 4. 我们提出的 AR-1-to-3 框架概览。左侧展示了 AR-1-to-3 的整体流程，右侧则展示了目标视角的去噪过程。以输入的单视图图像为起点，AR-1-to-3 利用扩散模型按从近到远的顺序逐步生成所有目标视角，前一步生成的视角作为物体本身的上下文信息参与后续生成。为实现该过程，我们设计了 Stacked-LE 与 LSTM-GE 两种策略，分别用于编码部分视角序列的局部特征与全局特征，作为去噪 UNet 的图像条件，辅助当前步骤的视角预测。

数量与特征维度。最终计算注意力输出为：

$$O_i = \text{Attention}([Q_i, K_i^*, V_i^*]). \tag{3}$$

3.4. 长短期全局特征编码

本节介绍如何将输入图像与已有目标视角序列的 CLIP 特征编码为全局条件，作为去噪 UNet 的交叉注意力输入，提供高层语义引导信息。我们在实验中观察到，条件视角序列的 CLIP 特征均为具有相同通道维度的一维向量。此外，这些特征可根据其对应视角的俯仰角划分为两个子序列，其方位角间隔为 120°。因此，使用长短期记忆网络（LSTM）[12] 处理此类序列是自然且合适的选择。

进一步而言，该策略具备以下两点优势，可显著减轻计算负担：1）无论参考视角数量多少，LSTM 可将其编码为两个向量；2）LSTM 结构具有良好的序列建模能力，且参数量较小。我们将该全局特征编码策略称为 LSTM-GE，其结构如 图 4 左下角所示。

在第 k 个自回归步骤中，存在 $2k-1$ 个条件视角。我们首先将这些图像输入 CLIP 模型的图像编码器，提取其视觉特征，记作 $F \in \mathbb{R}^{B\times(2k-1)\times D}$。随后，根据视角的俯仰角将这些特征划分为两组。注意，输入图像的特征作为特殊视角，会同时被包含在两个子组中。我们将两组特征分别记为 $F_1 \in \mathbb{R}^{B\times k\times D}$ 和 $F_2 \in \mathbb{R}^{B\times k\times D}$，并将其输入两个独立的 LSTM 模块。然后，从两个 LSTM 模块中提取第 k 步的隐藏状态，记为 h_l^k，作为各自的输出 I_l，其过程可表示为：

$$I_l = \text{LSTM}_l([F_l, (h_l^0, c_l^0)]), \quad l \in \{0, 1\} \tag{4}$$

其中，h_l^0 和 c_l^0 分别为隐藏状态和记忆单元，初始均为零向量。最终，将两个 LSTM 模块的输出在通道维度上拼接，并经过一层 MLP 与一组可训练的全局权重 $W \in \mathbb{R}^{77\times 1}$，得到用于去噪 UNet 交叉注意力的全局嵌入表示：

$$T = W \cdot \text{MLP}(\text{Concat}([I_0, I_1])), \quad T \in \mathbb{R}^{B\times 77\times D} \tag{5}$$

需要指出的是，我们在最终的全局条件中移除了 CLIP 的空文本嵌入，因为实验表明其对最终结果影响甚微。

4. 实验

4.1. 实验设置

数据集. 我们在最常用的三维基准数据集 Objaverse [7] 上进行实验，并引入两个域外数据集：Google Scanned Objects（GSO）[8] 和 OmniObject3D（Omni3D）[53]。

图 5. 我们提出的 AR-1-to-3 与当前主流多视角生成方法在新视角合成方面的可视化对比。与现有方法相比，AR-1-to-3 所生成的新视角在彼此之间以及与输入视角之间具有更高的一致性。

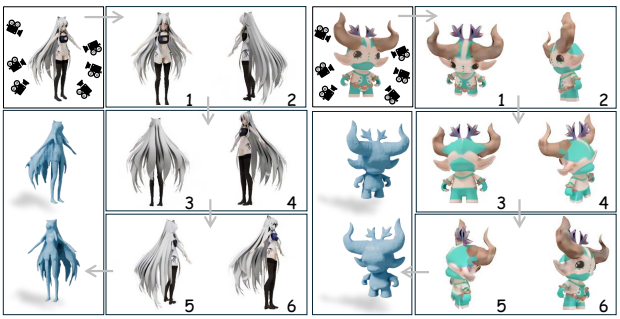

图 6. 基于我们的下一视角预测策略的图像到三维生成示例。AR-1-to-3 能够生成与输入图像一致的多视角图像，从而获得高质量的三维重建结果

参照先前工作的筛选策略 [18, 58]，我们从 Objaverse 数据集中筛选出约 21 万个几何物体，其三维网格容量为 80 万。我们从中选取 300 个涵盖多种类别的物体用于方法性能评估，其余样本用于训练。此外，为确保公平评估，我们还从 GSO 和 Omni3D 中分别随机选取 300 个样本。

按照 Zero123++ [35] 的实验设置，我们为每个物体渲染 7 张图像，包括 1 张输入图像和 6 张目标图像。具体而言，输入图像的相机视角通过在俯仰角 $-20°$ 到 $45°$、方位角 $0°$ 到 $360°$ 范围内随机采样获得。6 张目标图像的相机姿态由交替的绝对俯仰角（$20°$ 与 $-10°$）组成，其方位角相对于输入图像从 $30°$ 开始，每次递增 $60°$。此外，所有渲染图像的背景均设为白色，以确保扩散模型生成同类风格的图像，从而避免在三维重建过程中额外进行背景去除操作。我们将在项目中开源这些渲染图像。

评估方式. 我们从两个关键维度对各方法的性能进行评估，即二维图像保真度与三维几何准确性。具体而言，我们将多视角扩散模型生成的新视角图像或从合成的三维网格中渲染得到的图像，与真实视角图像进行比较。参考已有图像比较方法 [4, 20, 54]，我们采用了四个常用的评估指标：峰值信噪比（PSNR）、感知损失（LPIPS）[57]、结构相似性（SSIM）[51] 以及 CLIP 分数 [32]。我们还将从生成三维网格中随机采样的表面点与真实网格的表面点进行比较，采用的评估指标包括 Chamfer 距离（CD）以及阈值为 0.02 的 F-Score。

实现细节. 我们在 Objaverse 数据集中约 21 万个物体的渲染图像上对 AR-1-to-3 进行训练，总训练步数为 150k，使用 8 张 NVIDIA A100（80G）显卡，总 batch 大小为 32。学习率初始为 $1e$-5，每隔 25k 步进行一次周期性变化，优化器使用 AdamW [27]，调度策略为 CosineAnnealingWarmRestarts [26]。在训练过程中，我们从集合 $\{1, 2, 3\}$ 中随机采样一个 k，构建自回归模式，其中前 $2k-1$ 个视角作为条件图像，后两个视角作为目标图像。我们将条件图像的分辨率随机缩放至 128 至 512 范围内，以增强模型对不同输入分辨率的适应能力，并提高图像清晰度。同时，所有目标视角图像的尺寸统一调整为 320，因此自回归过程中生成的网格图像大小为 320×640。此外，我们采用了 Zero123++ [35] 中的线性噪声调度策略与 v-prediction 损失函数，而非稳定扩散模型 [33] 中的默认设置。在

推理阶段，AR-1-to-3 以输入图像为起点，分三步生成全部目标视角，具体如 图 6 所示。

4.2. 定性结果

我们在多个不同类别的三维物体上，开展了新视角合成与图像到三维重建的定性实验。为了突出 AR-1-to-3 在上下文推理与零样本泛化方面的优势，我们选用了相对于三维样本正面视角存在一定偏移的输入图像。更多可视化结果见附录材料。

新视角合成. 图 5 展示了我们提出的 AR-1-to-3 与当前主流多视角生成方法的合成结果，包括 Zero123-XL [23]、SyncDreamer [24]、Zero123++ [35] 与 One-2-3-45 [22]。需要注意的是，Zero123-XL 是在 Objaverse-XL 数据集上预训练的 Zero123 增强版本，One-2-3-45 的开源实现亦在其第一阶段使用了该版本以生成 8 个视角图像。我们使用 One-2-3-45 中提供的俯仰角估计模块，对 Zero123-XL 和 SyncDreamer 所需的相机俯仰角进行估计。在这些一致性难以保持的复杂场景中，一些方法生成了多个不一致的新视角图像，如 Zero123++ 在长椅场景下的结果以及 Zero123-XL 在卡通人物上的表现。还有一些方法在生成过程中出现混淆，导致结果与输入图像存在显著差异，如 SyncDreamer 对长椅的预测结果，及 One-2-3-45 中四轮床的生成图像。相比之下，我们的 AR-1-to-3 能够有效捕捉三维物体的纹理细节，并合成一致性良好的多视角图像，这得益于其对上下文信息的充分利用。

图像到三维重建. 得益于更加一致的多视角图像合成，我们的 AR-1-to-3 能进一步生成高质量的三维物体，如图 6 所示。我们还将其与五种最新的图像到三维方法进行了对比，分别为：SyncDreamer [24]、InstantMesh [54]、One2345++ [21]、TripoSR [44] 和 Unique3D [52]。需要说明的是，在可视化比较中，每个方法生成的网格都展示了几何形态（左侧）与纹理渲染图（右侧）。如 图 7 所示，AR-1-to-3 能够在仅有有限输入视角信息的条件下，生成外观一致、几何合理的三维网格模型。然而，对于其他方法来说，达成这一点依然是极具挑战的任务。例如，InstantMesh 和 One2345++ 容易在桌子物体后方错误地生成一个额外的储物柜。我们推测，这种现象源于它们在新视角生成过程中过度依赖扩散模型中的对称性先验，而对物体本身的上下文信息考虑不足。虽然 SyncDreamer 没有生成多余结构，但其重建出的桌子出现了显著的几何形变。我们发现，其原因在于 SyncDreamer 使用的扩散模型生成的 16 个视角之间存在较大的不一致性，从而影响了后续的三维重建。与上述方法不同，AR-1-to-3 在自回归生成所有目标视角的过程中，能够充分利用物体本身从近到远的上下文信息。因此，我们的方法在图像到三维的生成任务中展现出了优异性能。

4.3. 定量结果

我们在两个域外数据集（即 GSO 和 Omni3D）上进行了定量实验，以公平评估我们提出的 AR-1-to-3 与其他现有先进方法之间的性能差异。具体而言，所有的方法均接收相同的输入图像以生成三维资产。对于二维评估，我们对每个生成网格渲染 20 张分辨率为 224×224 的视角图像。在三维评估方面，我们从每个网格的表面均匀采样 16K 个点，采样范围为对齐到 $[-1,1]^3$ 的立方体坐标系中。如 表 1 所示，我们的方法 AR-1-to-3 在所有评估指标上均优于其他方法。这些结果进一步验证了 AR-1-to-3 在三维资产生成任务中的优越性。

4.4. 消融实验

我们在 300 个未用于训练的物体上进行消融实验，以验证 AR-1-to-3 框架中关键设计的有效性。

图像条件模块的消融分析. 从 Zero123++ 的基线方法出发，我们首先引入 Stacked-LE 策略，然后单独引入 LSTM-GE 策略，最后将两种策略结合使用。如 表 **??** 所示，两种策略分别都能带来性能提升，而将二者结合后则在基线上实现了更显著的改进。该结果不仅验证了两种图像条件策略的有效性，也表明我们所提出的下一视角预测机制在提升新视角合成准确性方面具有关键作用。

不同视角顺序的消融分析. 我们将相对于输入视角的相机位置从近到远的排列方式定义为 “正常顺序”。此外，我们还构建了两种顺序变体：即 “反向顺序” 与 “随机顺序”，用于目标视角的生成。具体来说，反向顺序是指相机从远离输入视角的位置逐步移动到靠近的位置。而随机顺序则将 3×2 网格中的中间一行为首行，后续再依次排列其余两行。如 表 **??** 所示，正常顺序的表现最优，而随机顺序效果最差，这表明以序列方式建模目标视角是有效的。需要指出的是，反向顺序与正常顺序表现接近。我们认为其原因在于：在 Zero123++ 的相

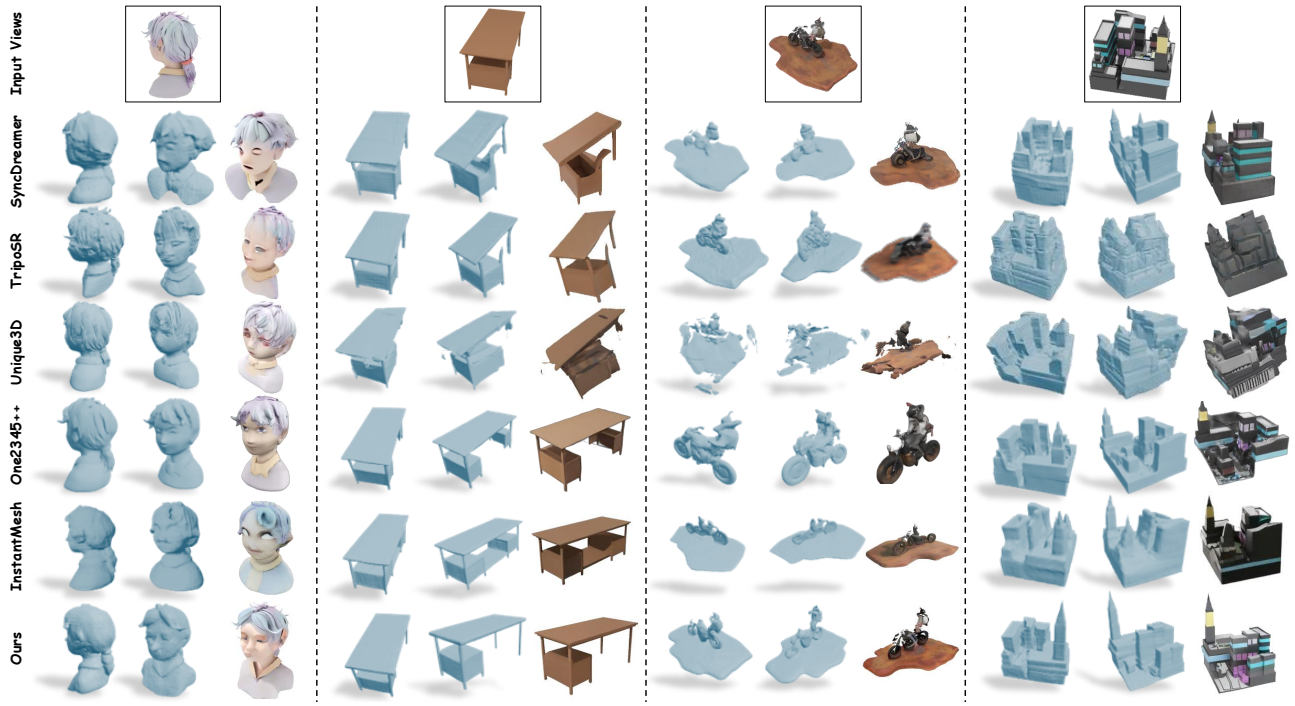

图 7. 我们提出的 AR-1-to-3 与当前先进方法在单视图图像生成三维物体任务中的可视化对比。需要注意的是，Unique3D、TripoSR 和 One2345++ 的三维重建结果是通过将输入图像输入其在 Huggingface 上的官方演示获得的。

表 1. 我们提出的 AR-1-to-3 模型与当前先进图像到三维方法的定量对比，涵盖三项二维视觉质量指标与三维几何质量指标。其中，“↑” 表示数值越高性能越好，“↓” 表示数值越低性能越好。

方法	GSO 数据集					Omni3D 数据集				
	PSNR ↑	SSIM ↑	LPIPS ↓	CD ↓	F-Score ↑	PSNR ↑	SSIM ↑	LPIPS ↓	CD ↓	F-Score ↑
Michelangelo [59]	9.323	0.609	0.408	0.165	0.105	9.969	0.602	0.410	0.174	0.081
SyncDreamer [24]	10.82	0.652	0.332	0.108	0.125	9.485	0.585	0.436	0.196	0.067
LGM [40]	9.139	0.592	0.429	0.157	0.075	10.02	0.588	0.394	0.152	0.086
InstantMesh [54]	10.67	0.661	0.338	0.117	0.135	9.91	0.608	0.412	0.178	0.076
AR-1-to-3 (Ours)	**13.18**	**0.709**	**0.232**	**0.063**	**0.258**	**10.25**	**0.629**	**0.388**	**0.148**	**0.097**

表 2. 针对 AR-1-to-3 条件组件的消融研究。

Stacked-LE	LSTM-GE	PSNR ↑	LPIPS ↓	SSIM ↑
		14.83	0.201	0.815
✔		17.59	0.174	0.833
	✔	17.91	0.170	0.836
✔	✔	20.28	0.121	0.857

表 3. 关于 AR-1-to-3 在自回归生成过程中视角顺序的消融研究。

视角顺序	PSNR ↑	LPIPS ↓	SSIM ↑	CLIP-Score ↑
反向	20.19	0.124	0.851	0.882
随机	17.36	0.167	0.839	0.774
正常 (Ours)	20.28	0.121	0.857	0.887

机设置下，反向顺序等价于相机以另一种方向沿三维物体进行圆周运动，即也是一种从近到远的环形路径。

全局特征序列的编码策略. 为验证 LSTM-GE 在视角序列全局特征编码中的有效性，我们设计了一种基于矩阵乘法的变体（简称 “matmul”），用于对这些特征进行编码。具体而言，我们首先将这些特征堆叠成一个形状为 $\mathbb{R}^{(2k-1)\times D}$ 的矩阵。同时，将全局条件中的可学习权重重复 $(2k-1)$ 次，得到一个形状为 $\mathbb{R}^{77\times(2k-1)}$ 的矩阵。随后，将这两个矩阵相乘，得到一个新的矩阵，其形状为 $\mathbb{R}^{77\times D}$，该矩阵被用作去噪 UNet 中交叉注意力机制的 key 和 value。如 图 8 所示，在引入

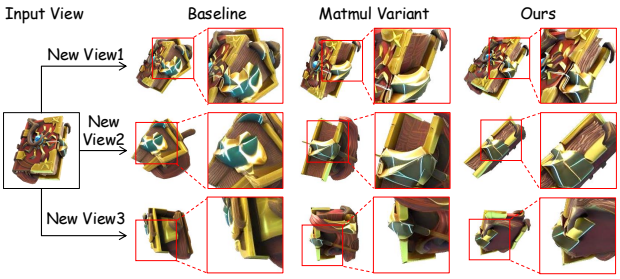

图 8. 关于全局特征编码策略的消融研究。

更多上下文信息后，该“matmul”变体相比基线方法（即 Zero123++）能够生成更加一致且质量更高的多视角图像。然而，该变体可能在对三维物体的全局语义理解上引入偏差，例如对书本样例的形状重建不准确。相比之下，我们提出的策略能够生成在形状和纹理上都更忠于输入视角的多视角图像。这些实验表明，采用 LSTM 的方案能够更有效地捕捉三维物体的高层语义信息。

5. 结论

本文提出了一种新的视角预测范式 AR-1-to-3，该方法从输入图像出发，逐步生成从近到远的目标视角图像。在每一步自回归过程中，先前生成的视角被用作上下文信息，以辅助当前目标视角的生成。实验结果表明，与现有的离散或同时生成多视角图像的方法相比，AR-1-to-3 能够生成在视觉外观与几何结构上更一致的新视角图像与三维物体。

References

[1] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2025. 3

[2] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22246–22256, 2023. 2

[3] Sijin Chen, Xin Chen, Anqi Pang, Xianfang Zeng, Wei Cheng, Yijun Fu, Fukun Yin, Yanru Wang, Zhibin Wang, Chi Zhang, et al. Meshxl: Neural coordinate field for generative 3d foundation models. *arXiv preprint arXiv:2405.20853*, 2024. 3

[4] Yabo Chen, Jiemin Fang, Yuyang Huang, Taoran Yi, Xiaopeng Zhang, Lingxi Xie, Xinggang Wang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Cascade-zero123: One image to highly consistent 3d with self-prompted nearby views. *arXiv preprint arXiv:2312.04424*, 2023. 1, 3, 6

[5] Yiwen Chen, Tong He, Di Huang, Weicai Ye, Sijin Chen, Jiaxiang Tang, Xin Chen, Zhongang Cai, Lei Yang, Gang Yu, et al. Meshanything: Artist-created mesh generation with autoregressive transformers. *arXiv preprint arXiv:2406.10163*, 2024. 3

[6] Yongwei Chen, Yushi Lan, Shangchen Zhou, Tengfei Wang, and XIngang Pan. Sar3d: Autoregressive 3d object generation and understanding via multi-scale 3d vqvae. *arXiv preprint arXiv:2411.16856*, 2024. 3

[7] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 5

[8] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 2, 5

[9] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3

[10] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. *arXiv preprint arXiv:2412.04431*, 2024. 3

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[12] S Hochreiter. Long short-term memory. *Neural Computation MIT-Press*, 1997. 3, 5

[13] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 4

[14] Hanzhe Hu, Zhizhuo Zhou, Varun Jampani, and Shubham Tulsiani. Mvd-fusion: Single-view 3d via depth-consistent multi-view generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9698–9707, 2024. 3

[15] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024. 3

[16] Black Forest Labs. Flux. *https://github.com/black-forest-labs/flux*, 2024. 2

[17] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023. 2, 3

[18] Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, et al. Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. *arXiv preprint arXiv:2502.06608*, 2025. 6

[19] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8640–8650, 2024. 1

[20] Yukang Lin, Haonan Han, Chaoqun Gong, Zunnan Xu, Yachao Zhang, and Xiu Li. Consistent123: One image to highly consistent 3d asset using case-aware diffusion priors. *arXiv preprint arXiv:2309.17261*, 2023. 3, 6

[21] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10072–10083, 2024. 2, 3, 7

[22] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3, 4, 7

[23] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 2, 3, 4, 7

[24] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 3, 7, 8

[25] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024. 1

[26] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6

[28] Yiwei Ma, Jiayi Ji, Xiaoshuai Sun, Yiyi Zhou, and Rongrong Ji. Towards local visual modeling for image captioning. *Pattern Recognition*, 138:109420, 2023. 3

[29] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 1, 2

[30] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2

[31] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 1, 3

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 2021. 4, 6

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 6

[34] David Ruhe, Jonathan Heek, Tim Salimans, and Emiel Hoogeboom. Rolling diffusion models. *arXiv preprint arXiv:2402.09470*, 2024. 3

[35] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 2, 3, 4, 6, 7

[36] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015. 3

[37] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint*, 2023. 3

[38] Yawar Siddiqui, Antonio Alliegro, Alexey Artemov, Tatiana Tommasi, Daniele Sirigatti, Vladislav Rosov, Angela Dai, and Matthias Nießner. Meshgpt: Generating triangle meshes with decoder-only transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19615–19625, 2024. 3

[39] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 3

[40] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 2, 8

[41] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *arXiv*, 2023. 3

[42] Zhenyu Tang, Junwu Zhang, Xinhua Cheng, Wangbo Yu, Chaoran Feng, Yatian Pang, Bin Lin, and Li Yuan. Cycle3d: High-quality and consistent image-to-3d generation via generation-reconstruction cycle. *arXiv preprint arXiv:2407.19548*, 2024. 3

[43] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*, 2024. 3

[44] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024. 7

[45] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pages 1747–1756. PMLR, 2016. 3

[46] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[47] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3

[48] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pages 439–457. Springer, 2024. 3

[49] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023. 2

[50] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023. 3, 4

[51] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[52] Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. *arXiv preprint arXiv:2405.20343*, 2024. 7

[53] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–814, 2023. 2, 5

[54] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 2, 3, 4, 6, 7, 8

[55] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2 (3):5, 2022. 3

[56] Lyumin Zhang. Reference-only control. *https://github.com/Mikubill/sd-webui-controlnet/discussions/1236*, 2023. 4

[57] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[58] Xuying Zhang, Yutong Liu, Yangguang Li, Renrui Zhang, Yufei Liu, Kai Wang, Wanli Ouyang, Zhiwei Xiong, Peng Gao, Qibin Hou, et al. Tar3d: Creating high-quality 3d assets via next-part prediction. *arXiv preprint arXiv:2412.16919*, 2024. 3, 6

[59] Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, Bin Fu, Tao Chen, Gang Yu, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. *Advances in neural information processing systems*, 36:73969–73982, 2023. 8

[60] Chuanxia Zheng and Andrea Vedaldi. Free3d: Consistent novel view synthesis without 3d representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9720–9731, 2024. 3

[61] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *arXiv preprint arXiv:2405.01434*, 2024. 1

[62] Qi Zuo, Xiaodong Gu, Lingteng Qiu, Yuan Dong, Zhengyi Zhao, Weihao Yuan, Rui Peng, Siyu Zhu, Zilong Dong, Liefeng Bo, and Qixing Huang. Videomv: Consistent multi-view generation based on large video generative model, 2024. 3