基于下一部分预测的高质量三维资产创造方法

张旭迎 [1*§], 刘宇桐 [2*], 李阳光 [34], 张仁瑞 [3], 刘雨霏 [5], 汪楷 [1],
欧阳万里 [35], 熊志伟 [2], 高鹏 [5], 侯淇彬 [1†], 程明明 [1]

[1] 南开大学 [2] 中国科学技术大学 [3] 香港中文大学 [4] 瓦嘶特 (VAST) [5] 上海人工智能实验室

项目主页：https://github.com/HVision-NKU/TAR3D

图 1. 所提出的 TAR3D 模型生成的 3D 资产库。我们采用 SyncMVD [39] 作为纹理合成器。

Abstract

我们提出了 *TAR3D*，这是一个全新的框架，由 *3D* 感知的向量量化变分自编码器（*VQ-VAE*）和生成式预训练 *Transformer*（*GPT*）组成，用于生成高质量的 *3D* 资产。本研究的核心思路是将下一个 *token* 预测范式所具备的多模态统一性及出色的学习能力，迁移应用到条件式 *3D* 物体生成任务中。为实现这一目标，*3D VQ-VAE* 首先将大量 *3D* 形状编码到一个紧凑的三平面潜空间中，并利用来自可训练码本的一组离散表征，在查询点占用率的监督下重建细粒度几何结构。随后，配备了名为 *TriPE* 的自定义三平面位置嵌入的 *3D GPT*，通过预填充提示标记以自回归方式预测码本索引序列，从而能够逐步对 *3D* 几何结构的构成进行建模。在多个公开 *3D* 数据集上进行的大量实验表明，在文本到 *3D* 和图像到 *3D* 任务中，*TAR3D* 能够取得优于现有方法的生成质量。代码和预训练权重将开源。

1. 引言

条件式 3D 物体生成旨在生成高质量的 3D 资产，这些资产在语义上与给定的提示（例如 2D 图像和文本）相符。随着基于扩散方法的重大进展，该领域取得了显著进步，其中包括分数蒸馏采样（SDS）优化 [7, 46]、多视图合成 [37, 52, 67] 以及 3D 感知扩散生成 [31, 66, 73]。然而，由于扩散模型与大型语言模型（LLMs）采用不同的范式，要构建一个跨多种模态的统一模型面临着巨大挑战 [1, 3, 12, 42]。

最近，诸如 MeshGPT [53] 和 MeshXL [8] 等开创性工作尝试将大型语言模型（LLM）的自回归方式引入 3D 网格生成中。然而，由于对网格面的顶点进行量

[*] 共同一作。
[†] 通讯作者。
[§] 这篇工作于上海人工智能实验室实习期间完成。

化，它们在处理长度为面数量 9 倍的序列时面临困难。这种过长的序列长度限制了它们在具有数十万个面的工业级 3D 资产中的应用。这促使我们研究是否有可能将具有任意数量面的网格编码为固定长度的序列，以减轻自回归建模的计算负担。

针对这一问题，关键在于对 3D 表示进行高效编码，使序列长度与面的数量无关。三平面表示法就具备这样的优良特性。与无序点云或冗余体素等其他 3D 表示方法不同，这种表示方法通过将 3D 信息压缩到三个固定尺寸的 2D 特征图中，能够在存储效率和强大的表达能力之间取得平衡。由于其高级的 2D 构成方式，它更适合离散化处理，并且还能借鉴图像量化方法中的策略，例如 LLamagen [55] 所采用的方法。基于三平面表示法，我们提出了 TAR3D，这是一种由 3D VQ-VAE 和 3D GPT 组成的新型 3D 自回归框架。

3D VQ-VAE 将 3D 形状编码为紧凑的三平面特征，利用一个可训练的、富含上下文信息的 3D 几何部件码本，从中获取一组离散嵌入。通过这种方式，无论面的数量多少，3D 网格都能被表示为具有三平面尺寸长度的特征序列，从而降低了对大量 GPU 资源的依赖。经过量化的潜在表征随后被解码为神经占用场，用于 3D 重建。为实现不同平面之间的信息交换，我们提出在解码过程中引入特征变形和注意力机制设计，以获取细粒度的几何细节。3D GPT 由预填充的提示嵌入驱动，对与量化三平面特征相对应的码本索引序列进行建模，从而能够以自回归方式实现条件式 3D 物体生成。为了保留更多空间信息，在生成过程中，我们还定制了一种名为 TriPE 的 3D 位置编码策略。在该策略中，每个平面的 2D 位置信息与三个平面相同位置之间的 1D 位置信息被有机融合。

为了验证我们所提出的 TAR3D 的有效性，我们在大量 3D 物体上开展了广泛实验。实验所用数据集包括两个常用的基准数据集，即 ShapeNet [6] 和 Objaverse [15]，以及一个域外数据集——Google Scanned Objects [19]。基于自回归方式以及我们精心设计的策略，TAR3D 能够生成高质量的 3D 资产，如图 Fig. 1 所示。定量和定性结果均表明，我们提出的 TAR3D 在性能上显著优于近期的前沿 3D 生成方法。

我们的贡献可总结如下：

- 我们提出了 TAR3D，这是一种由 3D VAE 和 3D GPT 组成的新型自回归框架，用于条件式 3D 物体生成。据我们所知，这是首次尝试利用三平面表示对 3D 物体进行量化，并逐部分生成高质量资产。
- 我们引入了特征变形和注意力机制设计，以捕捉细粒度的几何细节。
- 我们提出了一种 3D 位置编码策略，即 TriPE，以尽可能多地保留空间信息。

2. 相关工作

2.1. 3D 生成

早期的 3D 生成方法主要侧重于生成不同形式的 3D 模型，例如以文本 / 图像为条件或无条件的方式生成点云 [65]、网格 [9] 和体素 [5]。由于训练所用 3D 物体的类别和数量有限，这些方法往往泛化能力不佳。随着扩散模型在 2D 图像生成领域取得显著进展，大量研究人员开始探索将预训练的 2D 先验知识迁移到 3D 生成任务中。开创性的工作（如 DreamFusion [46]、SJC [60] 和 Fantasia3D [7]）将渲染视图输入预训练的 2D 扩散模型，并通过逐形状优化来进行知识蒸馏。尽管开启了一个新的时代，但这些方法存在一系列严重问题，例如耗时且多面性等。另一条研究路线的方法（如 Zero123 [37]、Zero123++ [52] 和 One2345 [35]）利用 3D 物体的渲染视图来微调预训练的扩散模型，以实现新视角或多视角生成。为了提高生成的多视图之间的一致性，Consistent123 [34] 和 Cascade-Zero123 [10] 引入了额外的先验信息，例如边界和冗余视图。SyncDreamer [38] 提出通过构建 3D 感知注意力机制，来关联不同视图之间的对应特征。这些方法可以结合稀疏视图重建模型（例如 NeRFs [43, 45, 61]、LRMs [23, 26, 67] 和 LGM [57]）来生成 3D 物体。然而，这些方法的间接生成方式可能会导致细节丢失或重建失败，因为它们严重依赖于多视图图像的保真度。

近来，一系列研究成果 [66, 71, 75] 涌现，它们借助 3D 感知扩散模型直接合成 3D 物体。具体而言，在这类方法中，3D 形状会被输入到预训练的 3D VAE 中，进而得到连续的潜在特征。与这些方法不同，我们的 TAR3D 避开了扩散机制，而是通过自回归生成离散的几何部分来创建 3D 物体。

2.2. VAE & VQ-VAE

Variational Autoencoder (VAE) [30] 变分自编码器（VAE）[30] 通常用于将高维输入信息映射到连续的

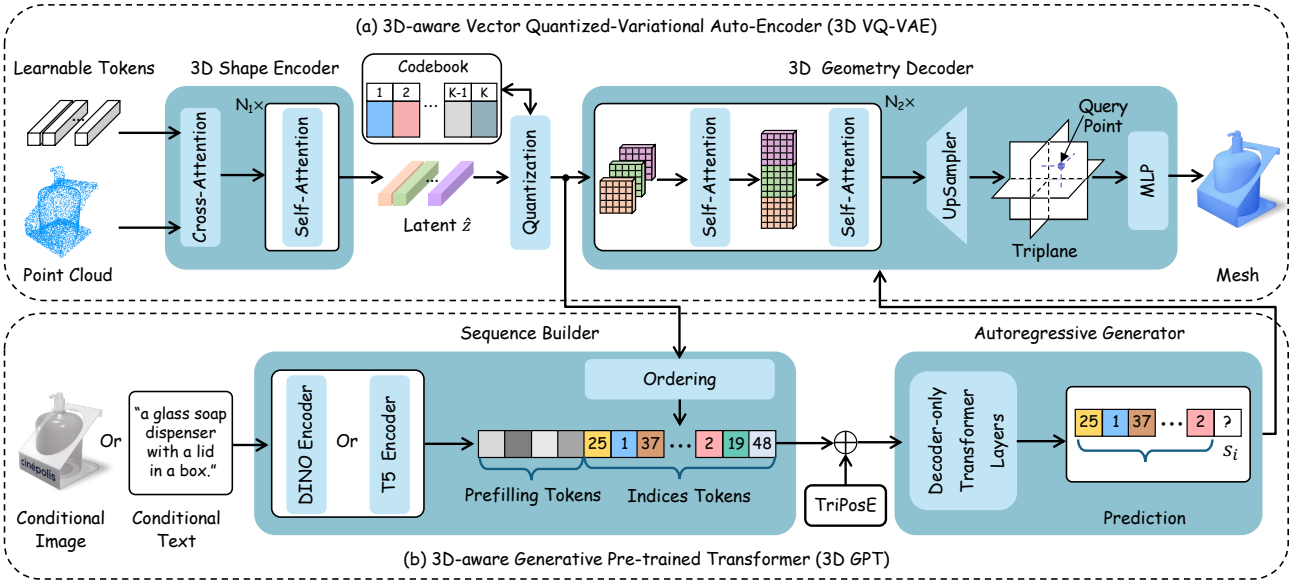

图 2. 所提出的 TAR3D 框架的整体架构。(a) 3D VQ-VAE 首先将从 3D 网格中均匀采样的点云编码为三平面潜在空间中的一组可学习 token。然后，这些连续的三平面特征被量化为来自可训练码本的离散嵌入向量。接下来，这些量化表示会经过两次变形，并在多个注意力层中配合两个自注意力模块，以实现每个平面内的特征增强以及三个平面之间的信息交互。随后，三平面特征被上采样至更高分辨率，以获取细粒度的几何细节。最后，从该三平面中采样的查询点特征被输入到一个 MLP 网络中，以进行占据状态预测。(b) 3D GPT 首先将来自 3D VQ-VAE 预训练码本的三平面索引组织成一个序列。在该序列中，每个平面内的索引按光栅扫描顺序排列，且三个平面中相同位置的索引按相邻顺序排列。然后，提示特征被用作该序列的预填充 token 嵌入，以实现条件性 3D 对象生成。接下来，这个序列通过仅解码器的多个 Transformer 层进行建模，采用下一部分预测的方式。通过查询码本，预测的索引序列可以被转换为三平面特征，从而合成 3D 对象。

概率潜在表示中，并且在生成建模领域具有深远影响。得益于这项工作，近期的扩散模型 [25, 50, 51] 能够在有限的计算资源上进行训练，同时保留其质量和灵活性。近期的 3D 生成方法，如 Clay [73]、Direct3D [66] 和 LN3Diff [31]，也基于该技术探索构建具有 3D 感知能力的扩散模型。向量量化变分自编码器（VQ-VAE）[59] 是变分自编码器（VAE）的一种变体。它引入了码本机制，将连续的潜在表示量化为离散组件，在许多生成任务上取得了良好的性能，例如文本到图像生成 [49]、音乐生成 [17, 18] 以及语音手势生成 [2]。VQ-GAN [20] 提出通过在训练过程中引入对抗损失来改进 VQ-VAE。在本文中，我们构建了一个具有 3D 感知能力的 VQ-VAE，用于对 3D 形状的三平面特征进行量化，并获取离散的几何部分，以用于自回归 3D 生成。

2.3. GPT

生成式预训练 Transformer（GPT）[47] 起源于自然语言处理（NLP）领域。基于仅解码器的 Transformer 架构，它按照下一个标记预测范式自回归地生成文本序列。这一系列具有突破性推理能力和惊人可扩展性的研究工作 [1, 3, 12, 42, 76] 不断涌现，为语言生成领域带来了变革。受这些成果的启发，许多研究人员已尝试将这一方案应用于图像生成领域。例如，Parti [69] 提出了一种路径自回归文本到图像模型，该模型将图像生成视为面向高保真图像的序列到序列建模。例如，LlamaGen [55] 证实，若进行适当的缩放，普通的自回归模型（如 Llama [58]）在不依赖视觉信号归纳偏置的情况下，也能实现最先进的图像生成性能。Emu3 [63] 通过将图像、文本和视频 token 化到离散空间中，在生成任务和感知任务上均取得了优异的性能。此外，AutoSDF [44] 学习了一种“非序列式”的自回归形状先验，用于 3D 补全、重建和生成任务。最近，一些研究工作如 MeshGPT [53]、MeshAnything [11] 和 MeshXL [8] 也尝试借助 GPT 以自回归的方式生成 3D 网格的面。

与这些方法不同，我们的 TAR3D 将 3D 形状的三平面表示量化为离散的几何部分，并利用 GPT 模型以

预测下一部分的方式生成 3D 对象。

3. 方法

Fig. 2展示了我们 TAR3D 框架的整体架构。我们的目标是将 GPT [1] 出色的学习能力和多模态统一能力迁移到条件 3D 对象生成任务中。然而，现有的对网格面进行量化的方法 [8, 11, 53] 存在序列过长的问题，这限制了它们在高质量 3D 资产生成中的应用。在这项研究中，我们提出使用三平面潜在表示来描述网格的 3D 形状信息，其特征图与三个轴平面（即 XY、YZ 和 XZ 平面）相关联。这些三平面特征可通过可训练的码本进行量化，从而形成一个固定长度的序列，而不受面数的影响。

3.1. 3D VQ-VAE

为了将 3D 形状的三平面表示量化为离散嵌入并合成 3D 对象，我们开发了一种 3D VQ-VAE，它包含 3D 形状编码器、量化器和 3D 几何解码器。

3D 形状编码器旨在获取 3D 对象紧凑且稳健的潜在表示，以保留细粒度的几何信息。对于一个 3D 对象，我们首先从其表面均匀采样高分辨率点云。为了增强点云的表达能力，点云表示中还包含了相应的法向量，记为 $\mathrm{P} \in \mathbb{R}^{B\times N_p\times(3+3)}$，其中 B 为批量大小，N_p 为点的数量。然后，我们对该点云表示应用傅里叶位置编码 [56]，以捕捉高频细节。

受先前点云理解工作 [28, 71] 的启发，我们采用了一种基于 Transformer 的架构，该架构由一个交叉注意力层和 N_1 个自注意力层组成，用于提取 3D 点云的潜在特征。具体来说，点云信息通过交叉注意力层被注入到一系列可学习的查询 token 中，这些查询 token 表示为 $\mathrm{e} \in \mathbb{R}^{B\times(3\times h\times w)\times d_e}$，其中 h 和 w 分别是三平面特征图的高度和宽度，d_e 表示这些可学习 token 的通道数。此后，这些 token 的表示能力通过后续的自注意力层得到增强，从而形成三平面潜在表示，即 $\hat{\mathrm{z}} \in \mathbb{R}^{B\times(3\times h\times w)\times d_z}$。

量化器的作用是使用来自可学习离散码本 $\mathcal{Z} = \{z_k\}_{k=1}^{K} \subset \mathbb{R}^{d_q}$ 的嵌入 z_q 来表示连续的三平面特征。具体而言，我们首先使用一个线性层将连续特征投影到与码本嵌入相同的通道数，得到特征 $\tilde{\mathrm{z}} \in \mathbb{R}^{B\times(3\times h\times w)\times d_q}$。然后，对这些特征的每个空间编码与其最接近的码本条目进行逐元素量化，具体如下：

$$z_q := \left(\underset{z_k\in\mathcal{Z}}{\arg\min} ||\tilde{\mathrm{z}}_{ij} - z_k|| \right) \in \mathbb{R}^{B\times(3\times h\times w)\times d_q} \quad (1)$$

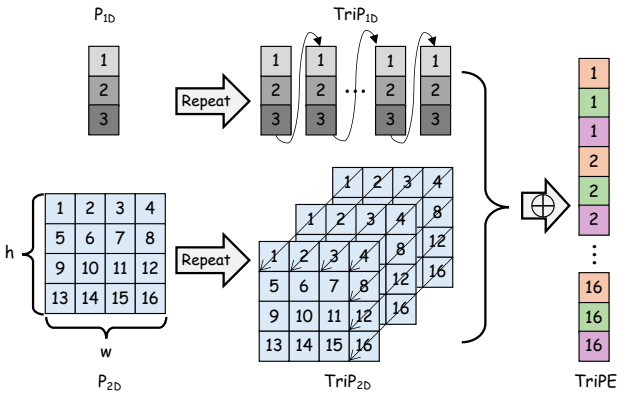

图 3. 我们为三平面序列的位置编码设计的 TriPE 的图示细节。我们用数字来表示位置信息，并且为了便于展示，简化了二维编码中的标记数量。

3D 几何解码器旨在以量化的离散特征（即 z_q）作为输入，高质量地重建 3D 神经场。受近期视频生成 [22, 27] 和虚拟形象生成 [62] 相关研究的启发，我们通过将 z_q 输入到由两个特征变形和自注意力操作构成的 N_2 个注意力层中，实现了平面信息的交互。具体而言，通过将平面轴重塑为批量轴来忽略平面轴，使得第一个自注意力能够独立处理每个平面。平面轴被恢复，且三个平面的特征沿高度维度进行拼接，这使得第二个自注意力能够对不同平面间的信息交互进行建模。我们还借鉴了 Direct3D [66] 中的操作，将三平面特征上采样至更高分辨率。对于 3D 场中由体素点和近表面点组成的一组查询点，可以通过双线性插值从生成的三平面中采样它们的特征。查询点特征通过多层感知器（MLP）转换为其占用值。

3.2. 3D GPT

为了以自回归的方式对 3D 对象的组成部分进行建模，我们提出了一种 3D GPT，它由三个组件构成，即序列构建器、TriPE 位置编码和自回归生成器。

序列构建器基于预训练的 3D VQ-VAE，3D 形状被编码为离散的三平面特征，这些特征可以用它们在码本中的索引来表示。随后，这些索引会依照某些排序规则被转换为一个序列。考虑到三平面表示由三个相互关联的特征图组成，我们按照光栅扫描顺序对每个平面内的索引进行组织，并将三个平面相同位置的索引按

相邻顺序排列。为实现条件性 3D 生成，提示词被编码为序列的预填充 token 嵌入。

TriPE 是一种为三平面索引序列量身定制的 3D 位置编码策略。如图 Fig. 3 所示，它是基于旋转位置嵌入（RoPE）[54] 的二维位置编码与一维位置编码的融合。我们将高度为 h、宽度为 w 的二维特征图的 RoPE 表示为 $\mathrm{P}_{\mathrm{2D}} \in \mathbb{R}^{h \cdot w}$，将包含 3 个标记的一维序列的 RoPE 表示为 $\mathrm{P}_{\mathrm{1D}} \in \mathbb{R}^{3}$。请注意，为便于描述，已移除通道维度。为了在三平面索引序列中保留轴对齐特征平面内的二维空间信息，我们将 P_{2D} 的单位元素重复三次，并将新生成的两个元素放置在其原始元素的相邻位置。我们将这种用于三平面索引的二维位置编码表示为 $\mathrm{TriP}_{\mathrm{2D}} \in \mathbb{R}^{3 \cdot h \cdot w}$。同时，我们将 P_{1D} 中的三个元素重复 $h \times w$ 次，以突出这三个特征图的差异，从而得到用于三平面索引的一维位置编码，记为 $\mathrm{TriP}_{\mathrm{1D}} \in \mathbb{R}^{3 \cdot h \cdot w}$。最后，我们通过对 $\mathrm{TriP}_{\mathrm{2D}}$ 和 $\mathrm{TriP}_{\mathrm{1D}}$ 进行逐元素相加来计算 TriPE。

自回归生成器。将三平面索引 token 化为序列 $s \in \{0, ..., K-1, K\}^{3 \cdot h \cdot w}$ 后，结合提示词 c 和自定义位置编码 TriPE，三维物体生成可表述为一个自回归的下一个索引预测问题。具体而言，解码器-Transformer 层学习预测可能的下一个索引的分布，这可以表示为：

$$p_\theta(s|c) = \prod_t p_\theta(s_t|s_{<t}, c), \tag{2}$$

其中，t 是生成过程中的时间步，c 是条件图像或文本嵌入，p_θ 表示带有参数 θ 的解码器-Transformer 层。

3.3. 优化细节

与 Fig. 2中的整体架构相对应，我们的 TAR3D 框架的优化过程也可分为两个阶段，即 3D VQ-VAE 优化和 3D GPT 优化。

为了以端到端的方式训练我们的 3D VQ-VAE，我们采用二元交叉熵（BCE）损失作为重建三维物体的优化目标。这一过程可形式化表示如下：

$$\mathcal{L}_{rec} = \mathbb{E}_{x \in \mathbb{R}^3}\left[\mathrm{BCE}\Big(\hat{\mathcal{O}}(x), \mathcal{O}(x)\Big)\right], \tag{3}$$

其中，$\hat{\mathcal{O}}(\cdot)$ 和 $\mathcal{O}(\cdot)$ 分别是查询点的预测占据值和真实占据值。对于量化器的码本学习，训练损失可以通过最小化原始特征与量化特征之间的差异来表示：

$$\mathcal{L}_{cb} = ||sg(\tilde{\mathrm{z}}) - z_q||_2^2 + \beta||\tilde{\mathrm{z}} - sg[z_q]||, \tag{4}$$

其中，$sg[\cdot]$ 表示停止梯度操作 [4]，β 是用于平衡两部分损失的权重超参数，默认设置为 $\beta = 0.25$。最后，我们通过最小化以下损失函数来优化 3D VQ-VAE：

$$\mathcal{L}_{3dvqvae} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{cb}\mathcal{L}_{cb}, \tag{5}$$

其中，λ_{rec} 和 λ_{cb} 分别是重建优化和码本优化的权重。

我们的 3D GPT 的优化目标是最大化三平面索引序列的对数似然。因此，其训练损失可表示如下：

$$\mathcal{L}_{3dgpt} = -\sum_{t=1}^{3 \cdot h \cdot w} \log\big(p_\theta(s_t|s_{<t}, c)\big). \tag{6}$$

4. 实验

4.1. 实验设置

数据集。为了验证我们的 TAR3D 的有效性，我们在两个基准 3D 数据集（即 ShapeNet [6]、Objaverse [15]）和一个域外数据集（即 Google 扫描物体数据集（GSO）[19]）上进行了实验。具体而言，ShapeNet 数据集提供了 52,472 个工业制造的网格模型，涵盖 55 个类别。我们采用了 3DILG [70] 的划分方式，其中 48,597 个样本用于训练，1,283 个用于验证，2,592 个用于测试。受先前数据过滤相关工作的启发 [32, 73]，我们对 Objaverse 数据集中 800,000 个网格模型的渲染法向图进行评分，得到了约 100,000 个几何物体。此外，我们随机选取了 1,000 个样本用于性能评估，其余样本则用于模型训练。此外，GSO 数据集包含约 1000 个真实世界的 3D 扫描模型，我们利用这些模型来进一步验证我们方法的泛化能力。对于这两个数据集中的每个 3D 资产，我们采用来自 ULIP [68] 的渲染图像和文本描述来构建提示系统。我们从所有渲染图像中均匀选取 4 张图像，并将其排名第 1 的描述文字用作文本提示。

评价指标我们从两个方面评估这些方法的性能：二维视觉质量和三维几何质量。在二维视觉评价方面，我们基于一系列常用指标，将从合成的三维网格模型中渲染出的新视角与真实视角进行比较，这些指标包括峰值信噪比（PSNR）、感知损失（LPIPS）[74]、结构相似性（SSIM）[64] 以及 CLIP[48] 分数。在三维几何评价方面，我们将从生成的网格模型和真实网格模型中随机采样得到的点簇进行比较。遵循先前研究中的方法 [36, 67]，我们采用倒角距离值和阈值为 0.02 的 F-Score 作为主要评价指标。

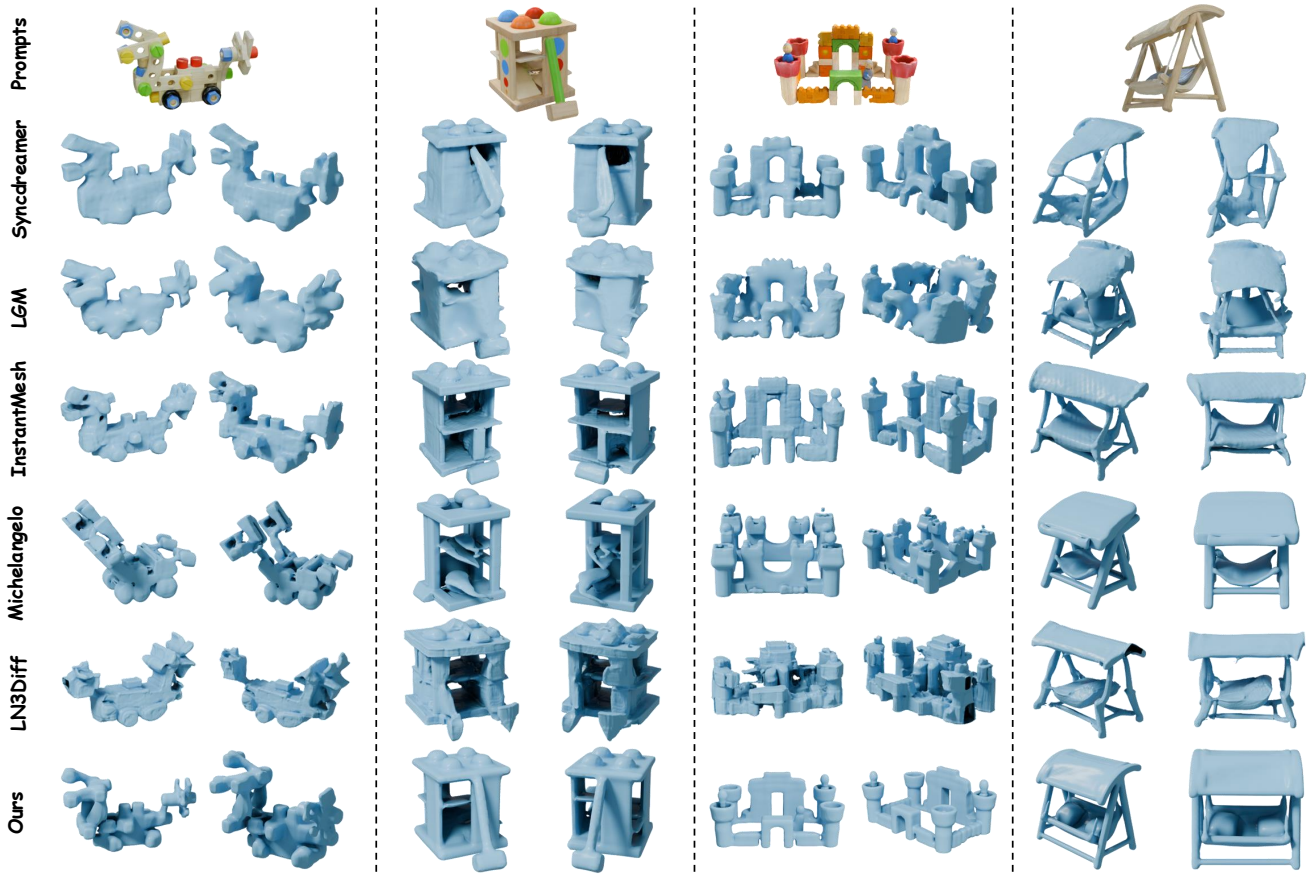

图 4. 我们将 TAR3D 与近期基于多视图的模型（即 Syncdreamer、LGM 和 InstantMesh）以及 3D 原生方法（即 Michelangelo 和 LN3DIff）在图像到 3D 对象生成任务中生成的 3D 网格进行了可视化比较。在给定来自 GSO 数据集的相同输入图像时，我们的 TAR3D 能够生成几何细节优于其他基线方法的 3D 资产。

实现细节。在我们的 3D VQ-VAE 中，输入到 3D 形状编码器的点云数量（即 N_p）为 81,920。3D 形状编码器中自注意力层的数量（即 N_1）为 8。三平面特征图的高度和宽度（即 h 和 w）均设置为 32。可学习 token 的通道数和三平面特征的通道数（即 d_e 和 d_z）分别设置为 768 和 16。码本的大小和通道数（即 K 和 d_q）分别设置为 16,384 和 8。此外，3D 几何解码器中注意力层的数量设置为 N_2=6。三平面特征被上采样至 256×256 的分辨率，其中 20,480 个均匀采样的体素点和 20,480 个近表面点被用作占用率监督的查询点。Eqn. (5) 中的超参数设置为 λ_{rec}=1 和 λ_{cb}=0.1。我们采用余弦退火调度器 [40]，将学习率初始化为 $1e{-}4$，并使其随时间逐渐衰减。我们采用 AdamW 优化器 [41] 来训练 3D VQ-VAE，在 8 块 NVIDIA A100 GPU 上以 128 的总批次大小训练 10 万步。

在我们的 3D GPT 中，我们分别采用预训练的 DINO 模型 [72]（ViT-B16 版本）和 FLAN-T5 XL 模型 [14] 对条件图像和文本提示进行编码。对于仅含解码器的 Transformer，我们遵循 LLamaGen [55] 的 GPT-L 配置，它包含 24 个 Transformer 层，注意力头数量为 16，维度为 1024。我们使用 AdamW 优化器训练 3D GPT，学习率设为 $1e{-}4$，总批量大小为 80，训练步数为 10 万。此外，在自回归推理过程中还引入了系数为 7.5 的无分类器引导（CFG）[24]，以提升几何质量和图像/文本-3D 对齐度。

4.2. 定性研究

图像到三维生成。我们首先展示了我们的 TAR3D 与近期用于图像到三维生成的最先进的模型之间的可视化对比，包括三种基于多视角的方法（如 SyncDreamer [38]、InstantMesh [67]、OpenLRM [23] 和 LGM [57]），以及三种原生三维方法（如 Shap-E [29]、

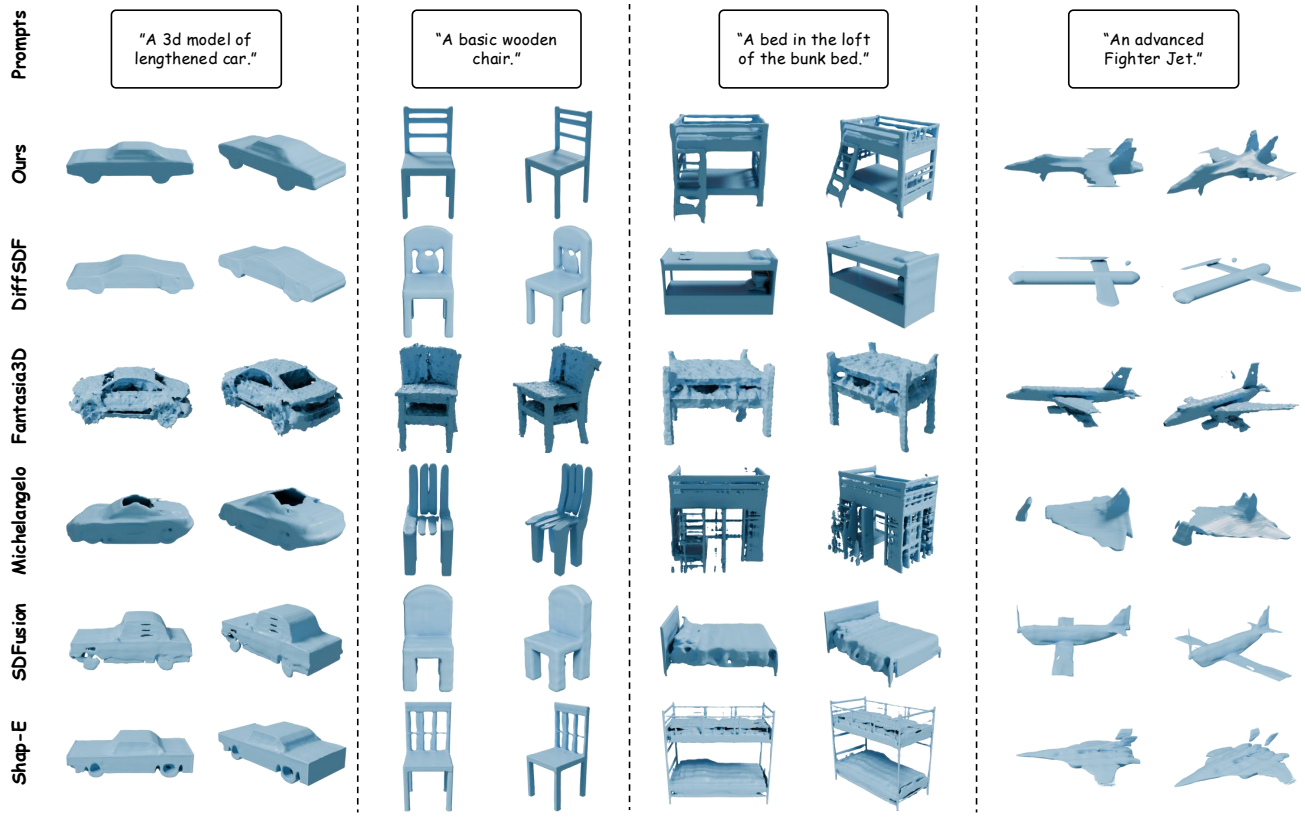

图 5. 我们将 TAR3D 与近期最先进的文本到 3D 对象生成方法进行了定性比较。与以往方法相比，我们的 TAR3D 所生成的 3D 网格资产在语义上与给定的文本提示更为契合。

Michelangelo [75] 和 LN3Diff [31])。

如 Fig. 4中的花园秋千样本所示，基于多视角合成的方法可能会生成具有不连续几何部分甚至错误的 3D 对象，这是由于它们用于重建的视角存在不一致性所致。同时，乐高样本显示，3D 扩散方法容易生成带有噪声的 3D 对象，这可能是由于条件去噪过程中存在异常所致。相比之下，我们的 TAR3D 模型凭借 GPT 架构强大的学习能力，能够合成具有卓越几何细节的高质量 3D 对象。

文本到三维生成。我们将我们的 TAR3D 与其他前沿的文本到三维生成方法进行了对比，包括 Diffusion-sdf [33]、SDFusion [13]、Shap-E [29]、Fantasia3D [7] 以及 Michelangelo [75]。如 Fig. 5所示，这些基准方法在多种情况下容易失效，要么生成质量不佳的 3D 对象（例如 Michelangelo 针对“一架先进战斗机”生成的结果），要么与给定的文本描述不匹配（例如 Fantasia3D 针对“双层床的上层床位”生成的结果）。与之不同的是，我们的 TAR3D 能够通过文本预填充嵌入驱动的 GPT 自回归过程，生成明显更合理的 3D 资产。

4.3. 定量评估

在本小节中，我们使用 ShapeNet 和 Objaverse 的混合评估集，从 2D 视觉质量和 3D 几何质量两个方面，将我们的 TAR3D 模型与近期的 3D 对象生成方法进行定量性能对比。考虑到从文本描述生成的 3D 对象具有多样性，我们在图像到 3D 任务上进行了实验以确保对比的准确性，如 Tab. 1所示。具体而言，所有候选方法均采用相同的图像作为条件输入来生成 3D 网格。在 2D 评估中，我们为每个网格渲染 20 个分辨率为 224×224 的视图，并将生成的法向图与对应真值视图的法向图进行比较。在 3D 评估中，我们在 $[-1, 1]^3$ 的对齐立方体坐标系中，从生成网格和真值网格的表面均匀采样 16K 个点云，以此对两者进行比较。我们的 TAR3D 模型在所有 2D 和 3D 指标上都大幅超越了其他前沿的 3D 对象生成方法。这些实验结果进一步证明了我们的 TAR3D 相较于现有方法的优越性。

表 1. 我们的 TAR3D 模型与近期最先进的图像到 3D 生成方法在 2D 视觉质量和 3D 几何质量上的定量比较。‘↑’：数值越高，性能越好；‘↓’：数值越低，性能越好。

Methods	Shap-E [29]	SyncDreamer [38]	Michelangelo [75]	InstantMesh [67]	LGM [57]	TAR3D (Ours)
PSNR ↑	10.991	11.269	11.928	11.560	11.363	13.626
SSIM ↑	0.702	0.706	0.734	0.721	0.714	0.763
Clip-Score ↑	0.834	0.837	0.864	0.847	0.841	0.868
LPIPS ↓	0.325	0.320	0.278	0.303	0.317	0.216
Chamfer Distance ↓	0.156	0.158	0.117	0.137	0.149	0.066
F-Score ↑	0.163	0.178	0.226	0.179	0.172	0.303

表 2. 关于 3D VQ-VAE 重建能力的消融实验。

	3D VAE	3D VQ-VAE
Chamfer Distance ↓	0.018	0.016
F-Score ↑	0.811	0.822

表 3. 关于 3D-VQVAE 中 PII 的消融实验。

	w/o PII	*w/* PII
Chamfer Distance ↓	0.023	0.016
F-Score ↑	0.661	0.822

表 4. 关于 3D GPT 的三平面尺寸（即序列长度）的消融实验。

Triplane Size	3×16×16	3×32×32	3×48×48
Chamfer Distance ↓	0.157	0.066	0.062
Inference Time ↓	17.7 s	67.6 s	143.9 s

4.4. 消融研究

3D VQ-VAE 的评估我们对 3D VQ-VAE 的 3D 重建能力进行了评估，该模型是生成高质量 3D 资产的基础。作为参考，我们提供了与我们的 3D VQ-VAE 同源的 VAE 模型的性能表现。为了获得一个适用于自回归生成的高性能分词器，我们对训练策略进行了调整，直至我们的 3D VQ-VAE 性能达到或超过同源 VAE 模型的水平。Tab. 2中的实验结果证明了我们的 3D VQ-VAE 具有良好的重建效果。

3D VQ-VAE 中的平面信息交互。我们分析了在 3D VQ-VAE 中，通过特征变形和注意力机制实现的平面信息交互（PII）的重要性。如 Tab. 3所示，未采用平面信息交互（PII）设计的变体在 3D 重建中获得了 0.661 的 F-score。通过融入我们的平面信息交互（PII）设计，该分数显著提升至 0.822。这些实验表明，我们的平面信息交互（PII）设计有助于从潜在特征到 3D 对象的解码过程。

TriPE 位置编码在 3D GPT 中的应用。我们针对 3D GPT 中的序列建模研究了两种位置编码策略。第一种策略是与序列长度相匹配的一维旋转位置嵌入（RoPE）[54]，另一种是我们提出的 TriPE。如 Fig. 6所示，一维旋转位置嵌入（RoPE）容易丢失输入图像中物体的重要几何细节。相比之下，我们精心设计的 TriPE 能够尽可能保留更多的 3D 空间信息，从而生成在几何上更接近提示文本的 3D 对象。

图 6. 关于我们所提出的 TriPE 有效性的消融实验。

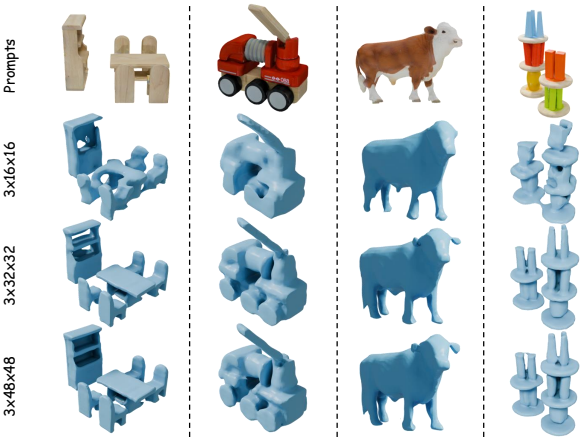

图 7. 我们的 TAR3D 变体在不同三平面尺寸或序列长度下的可视化比较。

三平面尺寸我们对三平面尺寸（3×16×16、3×32×32 和 3×48×48）进行了消融实验，以探究序列长度对性

能和效率的影响。需要注意的是，在 3D 重建中被广泛采用的 3×32×32 是我们的默认设置。如 Tab. 4所示，该设置在效率与性能的权衡中表现最佳，这一点也通过 Fig. 7中的视觉对比得到了印证。

5. 结论与未来展望

在本文中，我们提出了一个名为 TAR3D 的新型框架，该框架借鉴了多模态大型语言模型中的“下一个 token 预测”范式，用于生成高质量的 3D 资产。为实现这一目标，我们首先开发了一个 3D VQ-VAE 模型。在该模型中，3D 形状被编码到三平面潜在空间，并被量化为来自可训练码本的离散嵌入向量。借助精心设计的特征变形和注意力机制，这些量化特征被用于在神经占据场中重建细粒度几何结构。然后，我们采用这些离散表示的码本索引来构建用于自回归建模的序列。在预填充提示嵌入的驱动下，我们的 3D GPT 结合所提出的 TriPE 提供的 3D 空间信息，能够逐步生成高质量的 3D 对象。最后，我们在多个基准 3D 数据集和域外 3D 数据集上进行了大量实验，以证明与现有方法相比，我们的 TAR3D 具有更优异的性能。

我们认为，在自回归框架下改进 3D 对象生成具有广阔的前景。未来有几个值得探索的方向：1) 缩放定律：我们将从其他数据集（如 Objaverse-XL [16]、3D-FUTURE [21]）收集更多带提示的 3D 数据，并扩展 GPT 模型规模以提升生成能力。2) 序列构建：我们将探索更高效的三平面索引序列构建方式，以优化生成建模效率。3) 分词：我们计划为 3D 表示和提示嵌入建立统一的码本。

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 3, 4

[2] Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM Transactions on Graphics (TOG)*, 41(6):1–19, 2022. 3

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1, 3

[4] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013. 5

[5] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16123–16133, 2022. 2

[6] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2, 5

[7] 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. 1, 2, 7

[8] 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. 1, 3, 4

[9] Wenzheng Chen, Huan Ling, Jun Gao, Edward Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. *Advances in neural information processing systems*, 32, 2019. 2

[10] 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. 2

[11] 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, 4

[12] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 1, 3

[13] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4456–4465, 2023. 7

[14] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. 6

[15] 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

[16] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024. 9

[17] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020. 3

[18] Sander Dieleman, Aaron Van Den Oord, and Karen Simonyan. The challenge of realistic music generation: modelling raw audio at scale. *Advances in neural information processing systems*, 31, 2018. 3

[19] 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

[20] 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

[21] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129:3313–3337, 2021. 9

[22] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 4

[23] Zexin He and Tengfei Wang. Openlrm: Open-source large reconstruction models. https://github.com/3DTopia/OpenLRM, 2023. 2, 6

[24] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[26] 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. 2

[27] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 4

[28] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 4

[29] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 6, 7, 8

[30] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[31] Yushi Lan, Fangzhou Hong, Shuai Yang, Shangchen Zhou, Xuyi Meng, Bo Dai, Xingang Pan, and Chen Change Loy. Ln3diff: Scalable latent neural fields diffusion for speedy 3d generation. In *European Conference on Computer Vision*, pages 112–130. Springer, 2025. 1, 3, 7

[32] 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. 5

[33] Muheng Li, Yueqi Duan, Jie Zhou, and Jiwen Lu. Diffusion-sdf: Text-to-shape via voxelized diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12642–12651, 2023. 7

[34] 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. 2

[35] 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. 2

[36] Minghua Liu, Chong Zeng, Xinyue Wei, Ruoxi Shi, Linghao Chen, Chao Xu, Mengqi Zhang, Zhaoning Wang, Xiaoshuai Zhang, Isabella Liu, et al. Meshformer: High-quality mesh generation with 3d-guided reconstruction model. *arXiv preprint arXiv:2408.10198*, 2024. 5

[37] 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. 1, 2

[38] 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. 2, 6, 8

[39] Yuxin Liu, Minshan Xie, Hanyuan Liu, and Tien-Tsin Wong. Text-guided texturing by synchronized multi-view diffusion. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 1

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

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

[42] Gen Luo, Xue Yang, Wenhan Dou, Zhaokai Wang, Jifeng Dai, Yu Qiao, and Xizhou Zhu. Mono-internvl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training. *arXiv preprint arXiv:2410.08202*, 2024. 1, 3

[43] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106, 2021. 2

[44] Paritosh Mittal, Yen-Chi Cheng, Maneesh Singh, and Shubham Tulsiani. Autosdf: Shape priors for 3d completion, reconstruction and generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 306–315, 2022. 3

[45] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 2

[46] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1, 2

[47] Alec Radford. Improving language understanding by generative pre-training, 2018. 3

[48] 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. 5

[49] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 3

[50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[51] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 3

[52] 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. 1, 2

[53] 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. 1, 3, 4

[54] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 5, 8

[55] 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. 2, 3, 6

[56] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 2020. 4

[57] 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. *arXiv preprint arXiv:2402.05054*, 2024. 2, 6, 8

[58] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3

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

[60] 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

[61] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 2

[62] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4563–4573, 2023. 4

[63] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 3

[64] 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. 5

[65] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29, 2016. 2

[66] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. *arXiv preprint arXiv:2405.14832*, 2024. 1, 2, 3, 4

[67] 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. 1, 2, 5, 6, 8

[68] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1179–1189, 2023. 5

[69] 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

[70] Biao Zhang, Matthias Nießner, and Peter Wonka. 3dilg: Irregular latent grids for 3d generative modeling. *Advances in Neural Information Processing Systems*, 35:21871–21885, 2022. 5

[71] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–16, 2023. 2, 4

[72] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 6

[73] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024. 1, 3, 5

[74] 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. 5

[75] 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, 2024. 2, 7, 8

[76] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3