Puppeteer: 装配和动画您的 3D 模型

Chaoyue Song[1,2], Xiu Li[2], Fan Yang[1], Zhongcong Xu[2], Jiacheng Wei[1], Fayao Liu[3], Jiashi Feng[2], Guosheng Lin[†1], Jianfeng Zhang[†2]
[1]Nanyang Technological University [2]ByteDance Seed
[3]Institute for Infocomm Research, A*STAR
https://chaoyuesong.github.io/Puppeteer

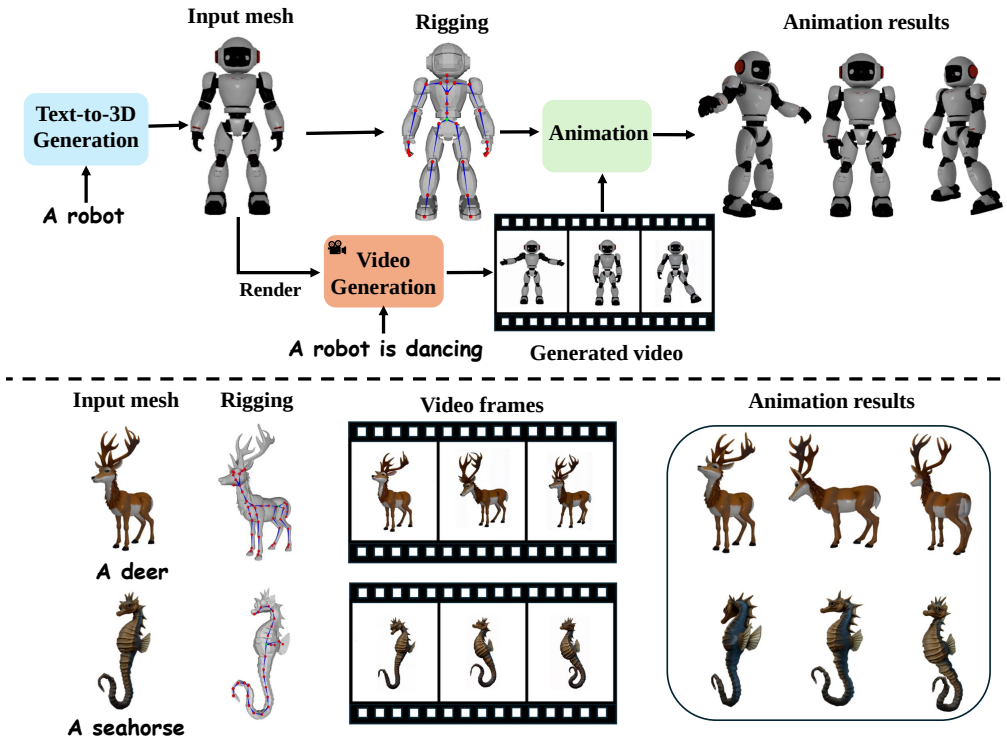

Figure 1: 给定一个三维模型，我们应用自动绑定来创建一个带有蒙皮权重的骨架结构。然后，将输入网格渲染为视频生成模型 [1, 34] 的输入。最后，我们根据生成的视频制作动画。输入的三维模型由 [73] 生成。

Abstract

现代交互式应用程序日益需要动态 3D 内容，但将静态 3D 模型转化为动画素材构成了内容创建流程中的重要瓶颈。尽管生成式 AI 的最新进展已在静态 3D 模型创建方面引发了革命，但绑定和动画仍然严重依赖专家的介入。我们提出 Puppeteer ，一个全面的框架，解决各种 3D 对象的自动绑定和动画问题。我们的系统首先通过自回归变压器预测合理的骨骼结构，变压器引入一种基于关节的分词策略用于紧凑表示，并采用具有随机扰动的分层排序方法以增强双向学习能力。然后，它通过一个基于注意力的架构来推断绑定权重，该架构包含拓扑感知的关节注意力，显式编码基于骨骼图距

† Corresponding authors. Email: chaoyue002@e.ntu.edu.sg.

Preprint. Under review.

离的关节间关系。最后，我们补充这些绑定方面的进展，通过一种基于可微分优化的动画管道来生成稳定且高保真的动画，同时在计算上比现有方法更高效。跨多个基准的广泛评估表明，我们的方法在骨骼预测准确性和蒙皮质量方面显著优于现有技术。该系统能稳健地处理多样化的 3D 内容，从专业设计的游戏资产到 AI 生成的形状，产生时间上连贯的动画，消除了现有方法中常见的抖动问题。

1 介绍

从 AAA 游戏和动画电影到 VR/AR 体验和机器人模拟，现代互动媒体需要动态的 3D 内容。尽管最近生成式 AI 的进步加速了具有复杂几何形状和纹理的高保真 3D 模型的创建，但这些资产仍然主要是静态的。将静态 3D 模型转换为动画版本需要两个专家驱动的过程：绑定（骨骼设置和蒙皮权重分配）和动画。这种手动且耗时的工作流程现在构成了现代内容创建流程效率的一个重大障碍。

研究界已投入大量精力自动化绑定过程。早期模板技术如 Pinocchio [6] 将预定义的骨架结构拟合到输入网格上，在特定类别上取得了令人满意的结果，但未能推广到任意形状。无模板算法 [25, 3, 7, 42, 71] 直接从几何属性中提取骨架结构，但常常产生过于密集或拓扑不兼容的关节配置，这不适合实际动画工作流程。深度学习方法已显著推进了这一领域：RigNet [85] 首创使用图神经网络从输入形状直接预测骨架和蒙皮权重，而 MagicArticulate [67] 将骨架生成重新表述为一个自回归问题，并引入了一个具有详细绑定注释的大规模数据集。尽管有这些创新，仍然存有重大挑战：RigNet 由于依赖精心设计的特征和严格的方向要求，在复杂网格拓扑中存在困难。MagicArticulate 在推断时计算效率低下，并且在蒙皮权重预测的功能扩散过程中推广性有限。关键是，这两种方法仅解决了管道中的绑定阶段，将同样具有挑战性的动画过程留作一个独立的手动任务，需大量的专业知识。

在这项工作中，我们提出了 Puppeteer，这是一个将自动绑定和动画集成到统一流程中的综合框架。为了应对现有数据集中的数据稀缺和姿势多样性有限的问题，我们将 Articulation-XL 数据集 [67] 扩展到 59.4k 个绑定模型，其中包括精心策划的 11.4k 多样姿势示例子集，以增强对不同姿势输入的泛化能力。这个扩展的数据集是我们基于学习的方法的基础。为了克服现有绑定方法在处理多样形状和复杂拓扑时的局限性，我们的系统对基础绑定组件进行了关键改进。在骨架生成方面，我们采用自回归变换器，具有基于关节的标记化和随机层次序列排序，创建更紧凑的表示，同时生成结构上连贯的骨架，不依赖模板。在蒙皮权重预测方面，我们提出了一种基于注意力的架构，结合了拓扑感知的关节点注意机制，显式编码骨骼图结构，实现了具有增强的泛化能力和计算效率的稳健权重预测。除了绑定之外，我们还解决了之前方法大多忽视的自动动画挑战。我们引入了一种基于可微优化的方法，无需神经网络参数，却通过结合我们生成的绑定与易于从现成视频生成模型中获取的参考视频指导，产生稳定的高质量动画。我们的统一框架实现了从静态网格到动画资产的全自动化，将劳动密集型的手动工作流程转变为多样 3D 内容创作的高效便捷流程。

广泛的评估展示了我们方法在绑定和动画任务中的有效性。对于绑定，在扩展的 Articulation-XL2.0 数据集和 ModelsResource 基准测试 [76, 84] 上的实验显示，在骨骼精度和蒙皮权重质量方面，相比于最先进的方法有显著的改进。我们的方法的稳健性通过成功应用于多样的 3D 内容得到了进一步验证——从专业设计的游戏资产到 AI 合成的几何体。对于动画，与最近的 4D 生成技术 [77, 59] 直接比较表明，我们基于优化的方法在保持计算效率的同时，产生了时间上一致性更高和视觉上更真实的结果。值得注意的是，我们的方法消除了在复杂运动序列中学习型方法常见的抖动伪影。干净且稳定的动画结果也突出了我们自动生成的绑定的可靠性。

总之，我们的工作通过四个关键贡献推动了自动化 3D 模型绑定和动画技术的发展：（1）扩展了包含 59.4k 绑定模型的大规模关节数据集，其中包括一个多样姿态的子集；（2）一种新颖的自回归骨架生成方法，该方法具有高效的基于关节的标记和随机策略的层次序列排序；（3）一种基于注意力的架构，用于包含拓扑感知关节注意力的蒙皮权重预测；以及（4）一种可微的基于优化的动画方法，能够为多种对象类别生成稳定且高质量的动画，而无需大量计算资源或手动操作。

2 相关工作

Skeleton generation. 生成三维模型骨架的方法主要分为两大类。第一类利用模板或额外输入。Pinocchio [6] 在自动骨架提取中开创了模板拟合，而 Li 等人 [37] 则利用深度学习结合给定的骨架模板来估计人体关节。一些最近的工作 [13, 24, 69] 继续沿用这一思路生成类人骨架。这些方法的一个显著局限是它们无法有效泛化到多样的对象类别。本组的其他方法则需要额外的输入，如点云序列 [86] 、网格序列 [14, 30] 、手动注释 [26] 、或视频数据 [87, 80, 66, 97, 89, 65, 68, 40] 。第二类方法则无需模板或注释。传统方法 [3, 7, 25, 71, 42] 提取曲线骨架，并且常常生成过于密集的关节点，不适合于动画制作。现代深度学习方法，如 Xu 等人 [84] 和 RigNet [85] ，直接从包含少于 3,000 个绑定模型的有限数据集中学习。尽管这些方法富有创新，但它们非常依赖精心设计的特征，并且在形状方向上施加的限制性假设极大地限制了它们在处理复杂网格拓扑时的有效性。

随着 3D 数据集 [15, 16] 的指数级增长以及自回归方法在 3D 生成 [62, 11, 12, 72] 中的成功，该领域在骨架生成方面取得了显著进展。MagicArticulate [67] 率先将骨架生成构建为一个自回归问题，并引入了一个包含绑定信息的大规模 3D 数据集 Articulation-XL。最近的一些工作 [45, 99] 也成功地结合了用于骨架生成的自回归变换器架构，进一步验证了这种方法。在我们的工作中，我们将 Articulation-XL 数据集大幅扩充，从 33k 扩大到 59.4k 个绑定模型，包括一个包含 11.4k 个例子的多样姿态子集。我们利用自回归变换器进行骨架生成，推出了两个关键创新：一种针对骨架结构的高效标记化方法，以及一种带有随机化的分层序列排序策略，以增强双向学习能力。

在骨架生成之后，自动绑定需要预测蒙皮权重以建立关节对网格顶点的影响。传统的几何方法 [18, 28, 19, 6] 根据顶点与关节的距离分配权重，这种方法对于复杂的拓扑结构而言是不够的。基于学习的方法 [46, 85, 54, 55] 通常结合图神经网络（GNN）和几何距离提示来进行蒙皮权重预测。然而，这些基于 GNN 的方法在可扩展性上存在显著局限性，并且难以在具有不同空间方向的 3D 数据上有效泛化。MagicArticulate [67] 将蒙皮权重预测表述为一个功能扩散问题 [96] ，但在推断速度和泛化能力方面表现欠佳。我们引入了一种基于注意力的网络，战略性地结合了骨架图距离，能够在不同对象类别中进行更为稳健的蒙皮权重预测，并显著增强泛化能力。并行研究 [99, 17] 同样利用表面点和骨骼之间的交叉注意来学习蒙皮权重。

有了生成的绑定，下一步是为 3D 模型制作动画。与早期专注于人类动作生成的工作 [31, 74, 100, 23, 101, 75, 63, 64, 60, 61] 不同，我们旨在为可以进行绑定的多样 3D 物体类别制作动画。我们的流程使用一个参考视频作为运动指导来为绑定的网格制作动画。4D 生成领域最近经历了快速增长，涵盖了文本/图像到 4D 生成 [90, 58, 44, 4, 50, 92, 8, 94, 41, 5, 70, 43, 83, 103] 和视频到 4D 生成 [21, 32, 95, 81, 97, 98, 39, 20, 105, 88, 82, 9, 59, 79] 。为获得全面的概述，我们建议读者参考 Miao 等人的综述 [51] 。然而，大多数现有的 4D 生成方法并不以特定 3D 对象为输入，以生成针对该对象的动画。

针对对象动画的多个重要尝试已经被提出。AnyMole [93] 引入了一种应用于各种类别的具有上下文动作的运动插补方法。AnyTop [22] 提出了在仅给定输入骨架的情况下，利用扩散模型无显性动作指导来动画化 3D 绑定网格的方案。Millan 等人 [52] 提出了相关工作，但专注于采用 SMPL [48] 代理的人形模型。Animate3D [33] 提出了使用多视角视频扩散模型来动画化 3D 对象，该方法需要大量训练。MotionDreamer [77] 尝试在不进行绑定的情况下动画化 3D 模型，但产生的动画质量不佳。最近的相关工作 AKD [38] ，使用一个 3D 模型，手动添加骨架，并使用 [6] 预测蒙皮权重。其后，使用基于视频的评分蒸馏采样 (SDS) 将动画应用于此绑定模型。然而，该方法计算密集（每个对象约需 25 小时）且产生不稳定的动画，伴随明显的抖动伪影。相比之下，我们提出了一种基于优化的方法，该方法不需要神经网络参数即可通过结合生成的绑定和参考视频指导来实现更稳定的多类别对象动画。

3 自动绑定

我们的自动绑定框架具有两个连续的模块。首先，我们部署一个自回归变压器，以从原始 3D 网格（Section 3.2 ）推断出结构上有效的骨架。随后，通过基于注意力的架构处理该骨架和原始网格，以预测精确的每顶点蒙皮权重（**??** ）。为了促进大规模学习，我们引入了 Articulation-XL2.0（Section 3.1 ），这是一个包含 59.4k 高质量绑定的 3D 模型的综合数据集。

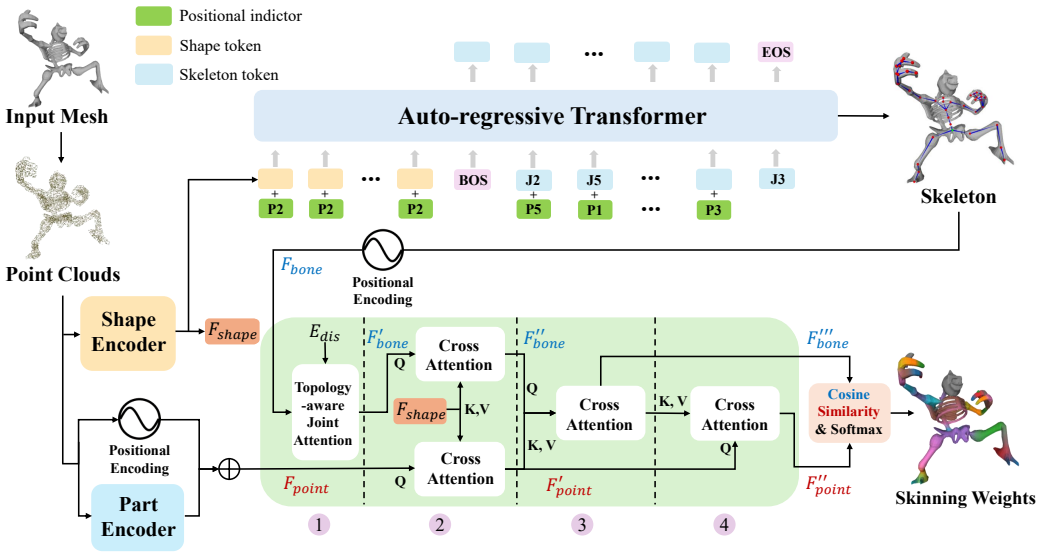

Figure 2: 我们的自动绑定流程概述。给定一个三维网格，我们首先采样具有法线的点云，然后使用自回归变压器生成骨架。点云和骨架通过一个基于注意力的网络进行处理，该网络包含四个关键操作： 1 通过拓扑感知的关节注意力增强骨骼特征， 2 通过与形状潜在特征的交叉注意力整合全局上下文， 3 通过交叉注意力实现骨骼与点的交互，以及 4 点特征优化。最后，余弦相似性和 softmax 归一化产生蒙皮权重。

3.1 数据集：Articulation-XL2.0

我们提出了 Articulation-XL2.0 ，这是在 [67] 中提出的 Articulation-XL 的扩展版本。我们的数据集结合了来自 Objaverse-XL [15, 16] 的多种几何数据类型，这些数据类型之前被排除在外，同时保持了相同的数据过滤过程。我们通过消除未蒙皮的顶点并进行人工验证来进一步提高质量，从而生成了超过 48k 个高质量的绑定 3D 模型。

意识到我们主要数据集中的模型主要处于休息姿态配置中，因此在推广到新姿态方面的能力有限，我们构建了一个多样姿态的子集。通过识别来自 Diffusion4D [41] 的高质量动画数据与我们的绑定模型语料库之间的交集，我们从动画帧中提取了 7.3k 变形网格，这些网格展现了与休息姿态配置的最大偏离，并附有相应的绑定信息。为了平衡该子集中类人形态的主导性，我们补充了使用 SMALR [106, 107] 生成的 4.1k 模型，这些模型的参数化基于 41 种不同动物的扫描和随机有效姿态生成。实验验证显示，最终得到的 11.4k 多样姿态数据集明显增强了对未见姿态的表现。我们将发布 Articulation-XL2.0 ，这是一个包含 59.4k 高质量绑定模型的综合集合，以促进未来的研究。数据集统计和示例在附录中提供。

3.2 自回归骨架生成

我们将骨架生成表述为一个形状条件序列建模问题。给定一个输入网格 $\mathcal{M}$ ，我们采用一个自回归框架（Figure 2 顶部）来预测一个骨架 $\mathcal{S}$ ，该骨架由 3D 关节位置 $\boldsymbol{J} \in \mathbb{R}^{j\times 3}$ 和由关节索引定义的拓扑骨连接 $\boldsymbol{B} \in \mathbb{N}^{b\times 2}$ 组成。我们的框架由三个关键组件构成：基于关节的骨架标记化、带有随机化的层次序列排序，以及形状条件自回归生成。这些组件共同作用，实现了对各种对象结构的准确、高效的骨架生成，而无须依赖预定义模板。

在 [67] 中，骨架被编码为基于骨骼的序列：每个 b 骨骼贡献 6 个标记（其两个端点的 3D 坐标），产生的总序列长度为 $6b$ ，并在多个连接的骨骼中重复关节位置。受到 [45] 的启发，我们开发了一种基于关节的标记化策略，通过关节的 3D 坐标和父索引来表示每个 j 关节，产生长度为 $4j$ 的序列。由于树状结构的骨架满足 $j = b + 1$ ，这产生当 $j > 3$ 时的 $4j < 6b$ ，使基于关节的表示更加紧凑。与通过 MLP 将关节位置投射到高维特征空间的 [45] 不同，我们将归一化的关节坐标离散化到 128^3 网格中，并附加父索引，生成作为我们自回归转换器输入的离散化标记序列。在实践中，我们为根关节分配父索引 0 ，并将所有其他父索引偏移 $+1$ （在去标记化过程中减去 1 ）。

Sequence ordering. 尽管我们基于关节的标记化提供了一种紧凑的表示方法，令牌的顺序对骨架的一致性和模型性能有显著影响。对于使用关节位置和父索引的骨架建模，令牌可以通过空间排序（升序的 z-y-x 坐标，如在 [67] 中）或层级排序（骨架树结构的广度优先遍历）来排序。我们的实验表明，空间排序经常产生不连贯的骨架，因为在生成父节点之前生成的子关节会创建无效的父引用（比较见 **??** 和附录）。因此，我们采用层级排序，仅在同一层级内的关节之间应用空间排序。

此外，受 [91] 的启发，我们通过序列随机化来增强双向学习能力。我们将每个关节的 4 标记组合在一起，并随机打乱这些组，结合目标感知位置指示符 $\boldsymbol{P} = [\boldsymbol{p}_0, \boldsymbol{p}_1, ..., \boldsymbol{p}_{j-1}]$ 来指导生成过程。具体来说，关节组内的所有标记共享一个位置指示符，指示下一个将生成哪个关节。为了识别第一个关节组，我们还在骨架标记 $\boldsymbol{T}_{skel}$ 之前将位置指示符整合到形状标记 $\boldsymbol{T}_{shape}$ 中：

$$\boldsymbol{T} = [\boldsymbol{T}_{shape}, \boldsymbol{T}_{skel}] + \boldsymbol{P} = [\boldsymbol{T}_{shape} + \boldsymbol{p}_0, \boldsymbol{T}^0_{skel} + \boldsymbol{p}_1, ..., \boldsymbol{T}^{j-2}_{skel} + \boldsymbol{p}_{j-1}, \boldsymbol{T}^{j-1}_{skel}]. \tag{1}$$

通过我们的分词策略和序列顺序建立后，我们现在描述自回归生成过程。我们从输入网格中采样 8,192 个带法线的点作为形状条件并使用预训练的形状编码器 [104] 对其进行编码。这个固定长度的形状标记序列 $\boldsymbol{T}_{shape}$ 位于 transformer 的骨骼序列之前，<bos> 和 <eos> 标记符号标记出骨骼的边界（在 Equation (1) 中被省略）。我们采用 OPT-350M [102] 作为我们仅使用解码器的 transformer 架构，使用交叉熵损失进行下一个标记预测的训练：

$$\mathcal{L}_{pred} = \text{CE}(\boldsymbol{T}, \hat{\boldsymbol{T}}), \tag{2}$$

其中 $\boldsymbol{T}$ 和 $\hat{\boldsymbol{T}}$ 分别代表真实值和预测的标记序列。在推理过程中，生成开始于形状标记和顺序位置指示符，自回归地进行，直到产生 <eos> 标记，然后解除标记以恢复完整的骨骼。

在本节中，我们介绍了一种基于注意力的网络，用于预测每个顶点的蒙皮权重，这些权重决定了网格在骨架关节运动时的变形方式。

Network architecture. 网络结构在 Figure 2 的底部进行了说明。我们的流程从输入网格中采样具有法线的 n 点开始。这些点经过位置编码和 PartField [47] 的部分编码器处理，获取结合了空间信息与部分特征的、具有部分感知能力的点嵌入 $\boldsymbol{F}_{point} \in \mathbb{R}^{n \times d}$ 。我们加入部分感知特征，因为部分与骨骼展现出强烈的解剖对应性，为蒙皮权重预测提供了有价值的结构指导。同时，通过连接每个关节的父位置与其自身位置来构建基于骨骼的坐标 $\in \mathbb{R}^{j \times 6}$ —— 对于根关节，其位置被复制以填充两个坐标槽。这些骨骼坐标同样经过位置编码以产生骨骼嵌入 $\boldsymbol{F}_{bone} \in \mathbb{R}^{j \times d}$ 。此外，我们将采样的具有法线的点输入到一个预训练的形状编码器 [104] 中，以提取全局形状潜变量 $\boldsymbol{F}_{shape} \in \mathbb{R}^{257 \times d}$ 。

该架构然后执行一系列注意力操作：(1) 骨特征增强。我们首先使用基于拓扑结构感知的关节注意力在骨嵌入上应用自注意力，以获得增强的骨特征。(2) 全局上下文整合。跨注意力在全局形状潜在变量（作为上下文）与点和骨特征之间执行，生成更新的特征。(3) 骨-点交互。跨注意力使用更新后的骨特征作为查询，并使用作为键/值以产生细化的骨特征。(4) 点特征细化。最终的交叉注意力在细化后的骨特征（作为上下文）与点特征之间执行，生成最终点特征。最后，网络计算余弦相似度得分并应用 softmax 归一化以产生蒙皮权重：其中是一个可学习的缩放参数。我们在训练期间使用交叉熵损失来优化网络。

Topology-aware joint attention. 尽管基本架构提供了有效的权重预测，我们的实验表明明确地建模骨骼结构显著增强了性能。我们的消融研究表明使用基于骨骼的坐标 $\in \mathbb{R}^{j \times 6}$ 而不是关节坐标 $\in \mathbb{R}^{j \times 3}$ 大大提高了性能（见 **??** ），这突出了骨骼结构中关节之间关系的重要性。

为了进一步利用拓扑结构，我们提出了拓扑感知联合注意力（Topology-aware Joint Attention，TAJA），它通过源于骨骼图距离的相对位置编码来增强标准自注意力机制。为了实现 TAJA，我们首先从骨骼结构计算图距离矩阵 $\boldsymbol{D} \in \mathbb{R}^{j \times j}$ ，然后通过量化和投影操作将这些距离转换为连续嵌入，从而生成位置嵌入 $\boldsymbol{E}_{dis} \in \mathbb{R}^{j \times j \times h}$ ，其中 h 是注意力头的数量。然后注意力机制修改为：

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}, \boldsymbol{E}_{dis}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}} + \lambda \boldsymbol{E}_{dis}\right)\boldsymbol{V}, \tag{3}$$

，其中 λ 是一个可学习的缩放参数。这种方法明确地结合了关节间拓扑关系，改善了网络理解骨骼结构并生成更准确蒙皮权重的能力。

4 视频引导的三维动画

利用生成的骨架和蒙皮权重，我们将静态网格转换为可用于动画的资源。本节介绍我们基于优化的方法，用于在视频指导下自动为已绑定的 3D 模型制作动画。
我们的动画过程从将绑定的网格渲染为初始帧 $\boldsymbol{I}_0$ 开始。利用这个作为条件图像，我们利用最新的文本到视频生成模型 [34, 1] 来在创建合理运动序列的同时保持对象的身份。利用描述所需动画的文本提示，这些模型生成一个视频序列 $V = \{\boldsymbol{I}_0, \boldsymbol{I}_1, ..., \boldsymbol{I}_{n-1}\}$，包含 n 帧。给定这个参考视频序列 V，我们联合优化 3D 网格的每帧关节旋转和全局根部运动，以使生成的动画与生成的视频序列对齐。

Differentiable optimization framework. 对于除第一个帧之外的每一个帧 $i \in \{1, 2, ..., n-1\}$，我们同时优化根部运动参数 $(\boldsymbol{Q}^i_{root}, \boldsymbol{T}^i_{root})$ 和关节特定旋转 $Q^i_{joint} = \{\boldsymbol{Q}^i_0, \boldsymbol{Q}^i_1, ..., \boldsymbol{Q}^i_{j-1}\}$，其中 $\boldsymbol{Q} \in \mathbb{R}^4$ 表示作为单位四元数的旋转，$\boldsymbol{T} \in \mathbb{R}^3$ 表示平移。对于第一个帧（静止姿态），我们将所有变换初始化为单位四元数和零平移，并在优化过程中保持不变。所有后续帧在优化开始前都以类似方式初始化。我们的优化过程结合了渲染损失、跟踪损失和正则化项：为了计算渲染损失，我们利用 Pytorch3D [57] 的可微渲染来生成预测帧 $\boldsymbol{I}'_i$，并计算这些预测与相应参考视频帧之间的 RGB、掩膜、光流和深度差异。视频帧的光流和深度通过现成的方法 [10, 53] 提取。跟踪损失结合了一个二维关节跟踪项和一个二维顶点跟踪项，这些项利用 Cotracker3 [35] 在整个视频序列中跟踪选定的点。我们将优化后的三维关节和变形的网格顶点投影到二维空间，并最小化它们与相应跟踪的二维关键点之间的距离。为了应对遮挡挑战，我们为关节和顶点实现了可见性检测机制。对于关节，我们基于光线与网格的交点来定义可见性：如果从相机投射到关节的光线与网格表面只有一次交点，则认为关节是可见的。我们使用 libigl [29] 的 ray_mesh_intersect 函数来计算这些关节可见性掩膜。对于顶点可见性，我们利用 Pytorch3D 的光栅化输出来确定可见的表面点。这些从第一帧导出的可见性掩膜确保我们的跟踪损失在整个序列中基于初始可见性一致应用，从而防止由于参考姿势中被遮挡的元素而产生的优化伪影。我们进一步加入正则项以加强帧间的运动平滑性。所有损失组件的完整数学公式在附录中提供。

我们在 Section 3.1 中引入的 Articulation-XL2.0 数据集上训练我们的模型，该数据集中包含超过 48k 个高质量样本，由 [15, 16] 的主要子集 Objaverse-XL 的样本和 11.4k 个来自多样化姿势子集的样本组成。在模型训练中，我们使用主要子集中的超过 46k 个样本和多样化姿势子集中的 10.9k 个样本。在评估时，我们使用三个不同的测试集：Articulation-XL2.0 测试（从主要集合中选取的 2k 数据），ModelsResource 测试 [76, 85]（270 个正面朝前且直立的模型，与 Articulation-XL2.0 不重叠，能够评估跨数据集的泛化能力），以及专门选定用于评估模型在不同姿势下表现的多样化姿势子集中的 500 网格部分。

为了增强鲁棒性和泛化能力，我们应用几何数据增强（缩放、平移、旋转变换）和姿态增强——将训练样本与其真实骨架和蒙皮权重结合，以模拟不同的姿势。进一步的实现细节在附录中提供。

我们包含四种比较方法作为基准：Pinocchio [6]，它将预定义的骨架模板拟合到输入网格。RigNet [85]，一种基于学习的模型，采用图卷积推断关节位置。MagicArticulate [67]，一个用于骨架生成的自回归框架，以及同时使用自动回归变压器方法的 UniRig [99]。所有方法都在 Articulation-XL2.0 和 ModelsResource 测试集以及我们多样化姿态子集上进行评估。我们采用 [84, 85] 中的三个基于 Chamfer 距离的度量来评估骨架生成质量：CD-J2J（关节到关节）、CD-J2B（关节到骨骼）和 CD-B2B（骨骼到骨骼）。这些度量衡量生成的骨架与真实骨架之间的空间对齐，较低的值表示性能更好。

Comparison results. 的定性结果在 Figure 3 中展示，涵盖所有三个基准。RigNet 一直生成无效的骨架——其图卷积模型在我们这个具有高度多样化方向的大规模数据集上训练时未能很好收敛。UniRig 展现了缺失和错位的骨架，例如乌龟四肢和松鼠尾巴上缺失的骨骼以及人类手上错位的骨架，如黄色圆圈所示。MagicArticulate 在 Articulation-XL2.0 和 ModelsResource 上与参考骨架非常接近，但在细节上存在错误（例如乌龟四肢缺失的骨骼，松鼠尾巴与身体错误连接）并在多样姿势子集上退化，因为它只在主要静态姿势数据上训练而没有进行姿势增强。相反，我们的方法在三个基准上产生了准确且结构正确的骨架。重要的是，我们生成的骨架甚至可以修正艺术家创建的骨架中的遗漏，例如缺失的乌龟头部与身体连接。Table 1 报告了定量指标，我们在每个数据集和指标上持续优于所有基线。值得注意的是，训练过程中引入多样姿势子集导致了在多样姿势基准上的显著改善。

我们评估了我们方法在 Tripo2.0 和 Hunyuan3D 2.0 生成的 AI 网格上的泛化能力。如 **??** 中所示，我们将我们的方法与 MagicArticulate 进行比较。MagicArticulate 失去了细节（例如，第 3 和第 5 行中的机器人的手，第 4 行中的海豚-蜂鸟奇美拉的尾巴和翅膀，用黄色标记）并产

Table 1: 骨架生成的定量比较。我们使用 CD-J2J、CD-J2B 和 CD-B2B 三个基准评估每种方法——结果均以 10^{-2} 为单位报告。较低的值表示更好的对齐。* 表示在 Articulation-XL2.0 上训练的模型，包括多姿态子集；未标记的模型则不是在其上训练的。加粗和下划线的数字分别表示最佳和次佳结果。

Method	Articulation-XL2.0			ModelsResource			Diverse-pose		
	J2J ↓	J2B ↓	B2B ↓	J2J ↓	J2B ↓	B2B ↓	J2J ↓	J2B ↓	B2B ↓
Pinocchio	8.324	6.612	5.485	6.852	4.824	4.089	7.967	6.411	5.149
RigNet	7.618	6.076	5.279	7.223	5.987	4.329	7.751	6.392	5.713
MagicArti.	3.264	2.503	2.123	4.114	3.137	2.693	4.376	3.456	2.955
UniRig	3.305	2.611	2.180	3.964	3.021	2.570	3.252	2.569	2.077
Ours	3.033	2.300	1.923	<u>3.841</u>	<u>2.881</u>	<u>2.475</u>	<u>3.212</u>	<u>2.542</u>	<u>2.027</u>
Ours*	<u>3.109</u>	<u>2.370</u>	<u>1.983</u>	3.766	2.804	2.405	2.514	1.986	1.598

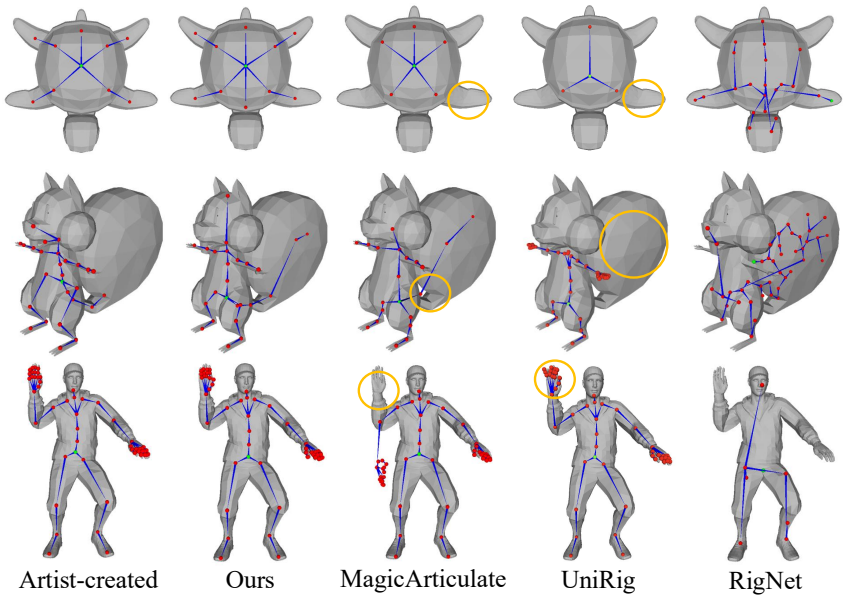

Figure 3: 定性骨架生成结果。数据来自 Articulation-XL2.0 、ModelsResource 和不同姿势子集，从上到下。

生了不对齐的骨架（第 1 行中的龙尾，第 2 行中的鹿腿）。相比之下，我们的方法在所有类别中一直生成有效且稳健的骨架。

我们将我们的蒙皮权重预测方法与三个基准进行比较：Geodesic Voxel Binding (GVB) [18] ，这是 Autodesk Maya [27] 中可用的几何基础技术，RigNet [85] 和 MagicArticulate [67] 。我们还在 Articulation-XL2.0 和 ModelsResource 测试集以及我们多样化姿势子集上评估这三种方法。蒙皮权重质量通过三个指标来评估：精确度、召回率和 L1 范数误差。精确度是预测权重 $>$ 1e$-$4 正确的比例，召回率是我们恢复的真实权重 $>$ 1e$-$4 的比例。L1 范数误差报告的是所有顶点的预测权重和真实权重之间的平均绝对偏差。变形误差结果在附录中提供。

Comparison results. Figure 4 通过其 L1 错误图可视化每个方法预测的蒙皮权重。我们的方法在所有基准上产生了更准确的权重分布，错误显著减少。RigNet 在所有示例上表现出较大的错误，而 MagicArticulate 的功能扩散在 Articulation-XL2.0 和多样化姿势子集上表现良好，但在 ModelsResource 上下降，揭示了有限的跨数据集泛化。在 Table 2 中的定量结果证实了这些观察结果，我们的方法在每个指标和数据集上都胜过所有基线。此外，我们的方法运行速度更快——实现的每个示例推理速度分别是 RigNet、MagicArticulate 和 GVB 的 1.75 × 、45 × 和 59 × （详情见附录）。

我们将我们的动画结果与 L4GM [59] 的视频到 4D 生成以及 MotionDreamer [77] 的 3D 网格动画进行比较。为了确保一个公平的评估，L4GM 被给予相同的输入视频，并且其第一帧的多视角合成被替换为输入 3D 模型的真实渲染。MotionDreamer 接收输入 3D 模型以及用于

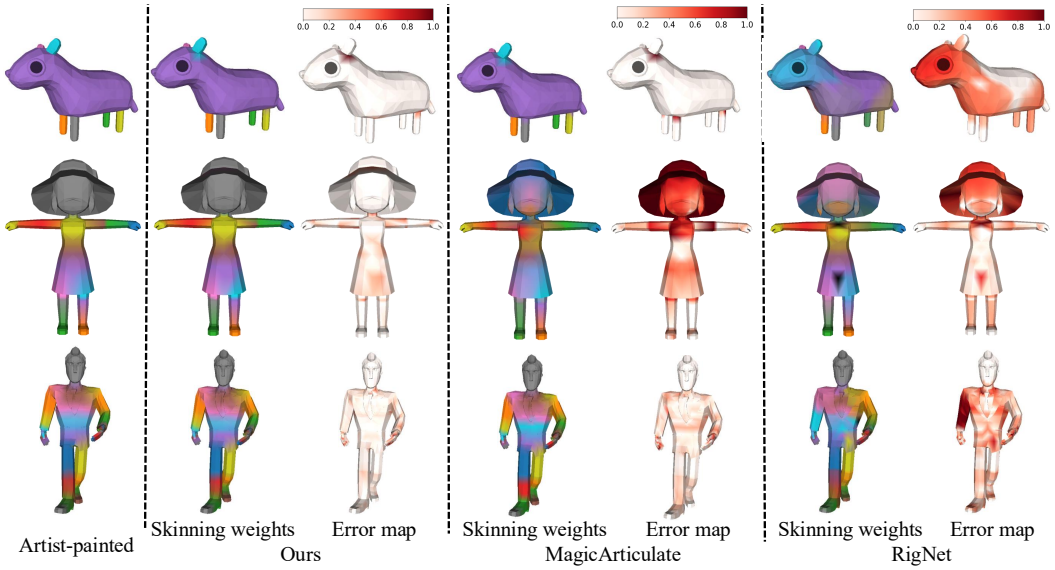

Figure 4: 定性蒙皮权重预测结果。数据来自 Articulation-XL2.0 、ModelsResource 和多样姿势子集，从上到下排列。每个示例都显示了预测权重的可视化及其 L1 误差图。附录中提供了更多结果。

Table 2: 蒙皮权重预测的定量比较。我们将我们的方法与 GVB、RigNet 和 MagicArticulate 进行评估。对于精确度（Prec.）和召回率（Rec.），数值越高表示更好的准确性和覆盖范围。对于平均 L1 范数误差（L1），数值越低越好。这里，* 表示在 Articulation-XL2.0 上使用多姿态子集训练的模型。

Method	Articulation-XL2.0			ModelsResource			Diverse-pose		
	Prec. ↑	Rec. ↑	L1 ↓	Prec. ↑	Rec. ↑	L1 ↓	Prec. ↑	Rec. ↑	L1 ↓
GVB	72.9 %	65.5 %	0.745	69.3 %	79.2 %	0.687	75.2 %	64.9 %	0.786
RigNet	73.7 %	66.1 %	0.729	65.7 %	80.2 %	0.707	74.7 %	65.4 %	0.746
MagicArti.	74.6 %	71.3 %	0.451	68.1 %	80.7 %	0.642	74.9 %	68.4 %	0.479
Ours	87.6 %	74.0 %	0.335	79.7 %	81.6 %	0.443	83.6 %	72.2 %	0.405
Ours*	87.9 %	73.8 %	0.333	79.8 %	81.5 %	0.442	86.4 %	72.8 %	0.353

视频生成的相同文本提示。在 Figure 5 中，其中一些输出看起来没有纹理，因为其密封网格转换破坏了 UV 映射。

如图 Figure 5 所示，我们展示了生成的骨架和相应的视频引导动画。带有骨架的形状代表静止姿态。虽然 L4GM 的参考视图与源视频对齐得很好，但它在提供地面真实多视图渲染时仍然反复产生几何失真（红色高亮）。MotionDreamer 的动画较为微妙，但可能会在刚性部分（例如，类人体躯干）中引入意想不到的变形。相比之下，我们的方法在使用完全生成的绑定时，能够生成准确且无伪影的动画。

在本节中，我们对骨架生成和绑定权重预测进行了消融研究。所有模型都在 Articulation-XL2.0 上训练，但不包括多样化姿势子集。

我们对四个组件进行了消融实验——姿态增强、顺序随机化、标记化方案和骨架排序策略——以衡量它们对骨架生成的影响（见 Table 3 ）。去除姿态增强会在所有基准测试中降低性能，尤其是在多样化姿态测试中。禁用顺序随机化同样会降低准确性。基于骨骼的标记化与我们的方法的质量相匹配，但需要额外 12 小时的训练时间，并且在推断时慢 1.6 倍。最后，用空间排序替换分层排序会保留 CD-J2J 和 CD-J2B，但明显增加了 CD-B2B 错误，并且经常产生断开的骨架；可视化比较请参见附录。

我们评估了蒙皮权重预测框架的四个关键组成部分（见 Table 4 ）。首先，用关节嵌入替换骨骼嵌入会使所有三个基准测试的平均 L1 范数误差增加 4.0 %，说明了显式建模骨骼信息的重要性。其次，将拓扑感知关节注意力（TAJA）替换为标准的自注意力会导致所有基准测试的性能下降，突出了建模关节之间拓扑关系的价值。第三，移除部件感知特征会导致性能

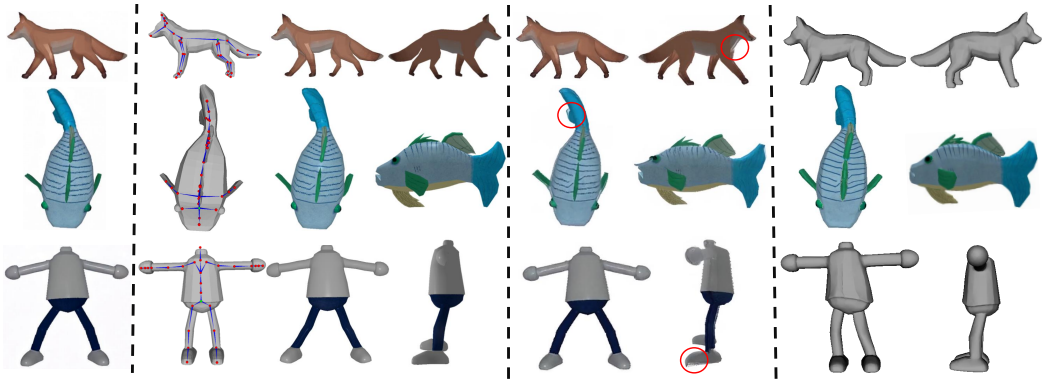

Figure 5: 动画结果比较。我们展示了生成的骨架和相应的视频引导动画。带有骨架的形状代表静止姿态。虽然 L4GM [59] 能够紧密地将其参考视图与输入视频对齐，但它始终表现出变形（用红色突出显示）。MotionDreamer 的 [77] 动画则显得微妙，并可能在刚性部分（例如，人形躯干）引入意外变形。相比之下，我们的方法通过完全生成的装配提供准确且无伪影的动画。视频已包含在项目页面中。

Table 3: 骨架生成的消融研究。

Method	Articulation-XL2.0			ModelsResource			Diverse-pose		
	J2J ↓	J2B ↓	B2B ↓	J2J ↓	J2B ↓	B2B ↓	J2J ↓	J2B ↓	B2B ↓
w/o pose aug.	3.131	2.451	2.223	3.994	3.141	2.843	4.886	4.029	3.629
w/o random	3.166	2.431	2.057	3.902	3.006	2.695	3.356	2.631	2.201
Bone token	3.014	2.309	1.939	3.865	2.940	2.524	3.269	2.518	2.087
Spatial order	2.982	2.298	2.068	3.868	2.961	2.641	3.210	2.570	2.295
Ours	3.033	2.300	1.923	3.841	2.881	2.475	3.212	2.542	2.027

持续下降，确认了它们在准确权重预测中的贡献。最后，在训练中去除姿势增强会导致在多样化姿势子集上的 L1 范数误差增加 9.6 %，表明姿势变化对于泛化到新姿势是必不可少的。这些发现确认了每个组成部分对我们模型整体准确性的重要性。

在这项工作中，我们引入了 Puppeteer，一个基于 59,400 个高质量绑定模型构建的统一绑定和动画管道。Puppeteer 首先使用基于关节的标记和包含随机化的层次排序的自回归转换器生成骨架。然后，一个基于注意力的网络利用拓扑感知特征预测蒙皮权重，随后一个高效的优化模块在低计算成本下生成稳定且高质量的动画。在多个基准测试中，Puppeteer 在骨架保真度、蒙皮准确性和动画流畅度方面优于最先进的方法。

5

致谢 本研究由教育部 AcRF 第二类基金（MOE-T2EP20223-0001）和教育部 AcRF 第一类基金（RG14/22）资助。

Table 4: 关于蒙皮权重预测的消融研究。

Method	Articulation-XL2.0			ModelsResource			Diverse-pose		
	Prec. ↑	Rec. ↑	L1 ↓	Prec. ↑	Rec. ↑	L1 ↓	Prec. ↑	Rec. ↑	L1 ↓
Joint embed	87.1 %	73.8 %	0.346	79.2 %	80.8 %	0.458	82.8 %	72.2 %	0.427
w/o TAJA	86.6 %	74.2 %	0.348	79.2 %	81.4 %	0.450	82.8 %	72.5 %	0.414
w/o part feat.	87.4 %	73.8 %	0.338	79.1 %	81.0 %	0.451	83.2 %	72.1 %	0.414
w/o pose aug.	88.0 %	73.3 %	0.337	79.3 %	80.7 %	0.449	82.8 %	70.1 %	0.444
Ours	87.6 %	74.0 %	0.335	79.7 %	81.6 %	0.443	83.6 %	72.2 %	0.405

在本附录中，我们提供了主论文的额外细节和实验结果，包括：

- Puppeteer （Appendix A ）和 Articulation-XL2.0 （Appendix B ）的进一步细节；
- 在骨架生成、蒙皮权重预测和动画方面的附加实验结果（Appendix C ）；
- 我们工作的局限性和未来工作的讨论（**??** ），以及更广泛的影响考虑（Appendix D ）。

A Puppeteer 的更多细节

我们的骨架生成首先通过预训练的形状编码器 [104] 对每个网格进行编码。我们首先使用 [78] 计算它的符号距离函数，然后使用 Marching Cubes [49] 重建粗略的网格，然后采样 8,192 个表面点（带有法线）。这些点最终被编码成固定序列的 257 个形状 tokens。我们将输入点归一化到 $[-0.5, 0.5]$ ，并对关节位置应用相同的缩放和位移进行对齐。然后将关节坐标离散化到一个 128^3 的网格上，添加父索引—生成长度为 $4j$ 的 token 序列。

在训练过程中，我们以 0.5 的概率应用姿态增强。当应用姿态增强时，每个关节有 0.3 的旋转概率，旋转角度限制在 $[-60°, 60°]$ 的范围内。序列排序的随机化遵循 [91] 进行退火，初始的置换概率为 r ，从 1 开始逐渐降到 0（恢复为层次顺序）在整个训练过程中：

$$r = \begin{cases} 1.0, & epoch \in [0, E/2], \\ 1 - \dfrac{epoch - E/2}{E/4}, & epoch \in [E/2, 3E/4], \\ 0.0, & epoch \in [3E/4, E], \end{cases} \tag{4}$$

其中，$epoch$ 是当前的训练阶段，而 E 是总的训练阶段数。此外，我们应用目标感知的位置信标来标记模型接下来应该生成的关节组。形状和骨架的标记序列（除了最后的关节组）都被增强了一个标识符，指示它们的下一个组；这些形状标记上的标识符特别用于标识要生成的初始关节组。自回归变压器在 8 个 NVIDIA A100 GPU（每个 GPU 的批量大小为 64，总有效批量为 512）上训练了大约 3 天 20 小时。

在训练过程中，我们基于注意力的网络条件设定对蒙皮权重进行预测，基于真实的骨架进行条件，并使用对应的蒙皮权重对其进行监督。我们从每个网格中采样 8,192 个表面点（带法线）——与我们的骨架流程匹配——并为每个点分配其最近顶点的权重。在推理过程中，这些预测的权重通过最近邻映射转回到网格顶点上。

训练在 Articulation-XL2.0 上进行，使用 8 个 NVIDIA A100 GPU，耗时大约 1 天 6 小时，每个 GPU 的批量大小为 16。一个有效关节掩码—支持多达 70 个关节—能够调整网络以适应在训练和评估过程中遇到的不同骨架大小。

Video-guided 3D animation. 在这里，我们详细描述了我们的视频引导 3D 动画过程。为了渲染监督，我们采用了四种不同的损失函数。给定生成的视频 $V = \{\boldsymbol{I}_0, \boldsymbol{I}_1, ..., \boldsymbol{I}_{n-1}\}$ 和使用 Pytorch3D [57] 渲染的图像 $\boldsymbol{I}'_i$, 我们计算以下渲染损失：

$$\begin{aligned} \mathcal{L}_{rgb} &= \sum_i \left\| \boldsymbol{M}_i \odot (\boldsymbol{I}_i - \boldsymbol{I}'_i) \right\|^2, & \mathcal{L}_{mask} &= \sum_i \mathrm{BCE}(\boldsymbol{M}_i, \boldsymbol{M}'_i), \\ \mathcal{L}_{depth} &= \sum_i \left\| \boldsymbol{M}_i \odot (\boldsymbol{D}_i - \boldsymbol{D}'_i) \right\|^2, & \mathcal{L}_{flow} &= \sum_i \left\| \boldsymbol{M}_i \odot (\boldsymbol{F}_i - \boldsymbol{F}'_i) \right\|^2, \end{aligned} \tag{5}$$

其中 $\boldsymbol{M}_i$ 表示视频帧中前景物体的二值掩码，$\boldsymbol{M}'_i$ 表示通过 Pytorch3D 渲染的 3D 物体的掩码。我们使用 [10] 提出的方法提取深度图 $\boldsymbol{D}_i$ ，而 $\boldsymbol{D}'_i$ 则直接从 Pytorch3D 渲染器中获得。我们应用尺度偏移对齐来处理相对深度 $\boldsymbol{D}_i$ 和度量深度 $\boldsymbol{D}'_i$ 之间的尺度模糊问题。对于光流估计，我们使用 [53] 的方法计算 $\boldsymbol{F}_i$ ，并通过将 3D 顶点流投影到二维图像平面上来导出 $\boldsymbol{F}'_i$ 。逐元素乘法运算符 $\odot$ 表示 $\mathcal{L}_{rgb}$ 、$\mathcal{L}_{flow}$ 和 $\mathcal{L}_{depth}$ 都只在由掩码 $\boldsymbol{M}_i$ 定义的前景区域内计算，确保我们的优化集中在目标物体上。

追踪损失包括一个二维关节追踪项和一个二维顶点追踪项，利用 Cotracker3 [35] 来追踪整个视频序列中的选定关键点。我们将三维静态物体的可见关节和顶点投影到图像平面上，以建立第一帧的关键点。这些关键点随后通过整个视频序列使用 Cotracker3 进行追踪，以获得

$\boldsymbol{p}_i$，代表每帧的追踪位置 i。同时，$\boldsymbol{p}'_i$ 通过将变形关节和网格顶点投影到图像平面上来获得。追踪损失被制定为：

其中 $\boldsymbol{M}_j$ 和 $\boldsymbol{M}_v$ 分别表示第一帧中关节和顶点的可见性掩码。这些掩码确保只有可见的关键点才对优化过程有贡献。此外，我们还结合了正则化项，通过惩罚连续帧之间的大幅变换变化来防止时间抖动。为了平衡这些损失，我们对每一项进行加权以确保其量级可比。在实际操作中，正则化损失相对于渲染和跟踪损失的权重被降低了 3-4 个数量级，以防止过度平滑。

我们的动画优化在单个 NVIDIA A100 GPU 上对于顶点数高达 10K 的对象大约需要 20 分钟，用于处理由 Kling AI [1] 或 JiMeng AI [34] 生成的 5 秒视频（大约 50 帧，10 FPS）。运行时间随着网格复杂性和帧数的增加而变化：（1）网格复杂性：具有更多顶点的模型将需要额外的 Pytorch3D 渲染时间。例如，我们项目页面中的蝙蝠案例（~ 为 70K 顶点）需要 90 分钟，而乌龟案例（~ 为 15K 顶点）需要 35 分钟。（2）帧数：对于一个典型的案例，在 50 帧（10 FPS，5 秒）的情况下需要 20 分钟，增加到 20 FPS（100 帧）将优化时间延长到 41 分钟，而减少到 4 FPS（20 帧）则将时间减少到 8 分钟，显示出近似线性的帧数缩放。

优化后，我们的动画过程遵循标准的骨骼动画原则 [56]：(1) 正向运动学（FK）：我们使用层次化的正向运动学从优化后的局部变换计算全局关节变换，从根节点遍历到叶节点；(2) 线性混合蒙皮（LBS）[36]：使用关节变换的加权组合来变形网格顶点，其中每个顶点根据蒙皮权重受到多个关节的影响。这产生了最终的网格动画序列。

A.1 实验细节

为了进行基线比较，我们使用了 UniRig [99]、RigNet [85] 和 Pinocchio [6] 在各自 GitHub 存储库中的公开实现。Geodesic Voxel Binding (GVB) [18] 的比较使用了 Autodesk Maya [27] 中的实现。RigNet 和 MagicArticulate [67] 在 Articulation-XL2.0 上进行了训练，并使用了作者指定的原始数据预处理流程和训练计划。

B Articulation-XL2.0 的更多细节

我们的数据集 Articulation-XL2.0 来源于 Objaverse-XL [15, 16]，特别关注包含绑定信息的文件类型的 GitHub 和 Sketchfab 子集（例如，glb/gltf，fbx，dae，blend 等）。从最初的 297 万个模型中，我们提取了 74K 个绑定资产（在删除了超过 15 万个重复项之后），并通过质量验证整理了 48K 个高质量的绑定模型。Table S 5 阐述了这些统计数据。Articulation-XL2.0 中的骨骼数量范围为 2 到 100，分布情况如 Figure 偏微分方程 6 所示。

Articulation-XL2.0 中的多样姿势子集来源于两个部分。第一部分由动画数据中提取的姿势组成，我们特别选择了与静止姿势配置最大偏差的帧，以捕捉极端的关节变化。第二部分包括使用 SMALR [106, 107] 合成生成的姿势，这些姿势基于 41 个不同动物扫描的参数化以及随机产生的有效姿势。这些随机有效姿势通过对 SMALR 动物关节应用随机旋转角度生成，同时限制角度在解剖学上有效的范围内，确保多样的关节状态并保持生物学合理性。请参阅 Figure 偏微分方程 8 以了解 SMALR 数据的骨骼结构和关节名称。最初在索引 0 中有两个关节，我们将它们合并成一个根关节。来自这个多样姿势集合的一些例子示例在 Figure 偏微分方程 7 中展示。注意，多样姿势测试集中模型及其相应的静止姿势完全排除在训练数据之外，以确保对新形状和新关节变化的泛化进行严格评估。

Table S 5：数据静态用于 Articulation-XL2.0 。

Source	All models	with rigging	high-quality rigging	low-quality rigging
Sketchfab	0.89M	64K	42K	22K
GitHub	2.08M	10K	6K	4K
Total	2.97M	74K	48K	26K

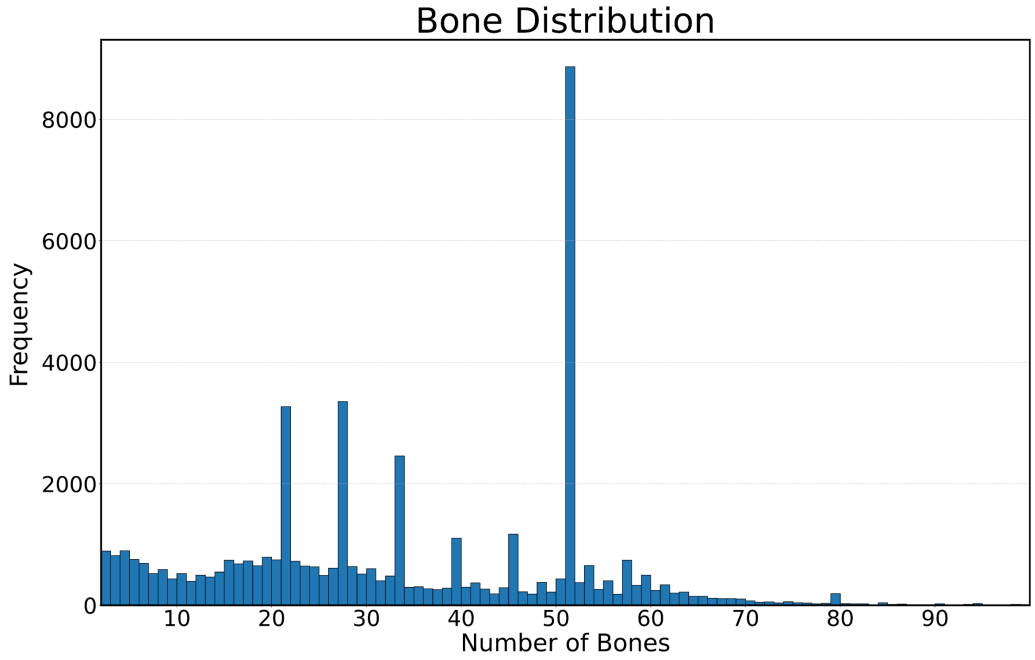

Figure 偏微分方程 6 : Articulation-XL2.0 的骨骼数量分布。

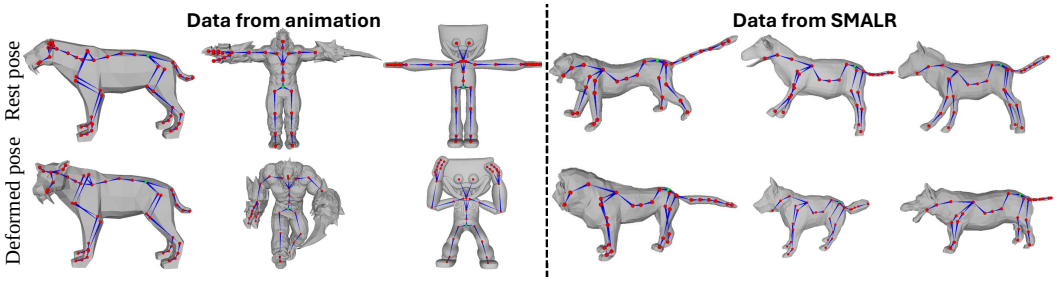

Figure 偏微分方程 7 : 来自 Articulation-XL2.0 的多样姿势子集的例子。左边的组展示了由动画衍生的样本：顶部显示的是原始的静止姿势，而底部显示的是从动画序列中在最大姿势偏差的帧处提取的变形关节。右边的组展示了使用 SMALR [106, 107] 合成生成的关节。

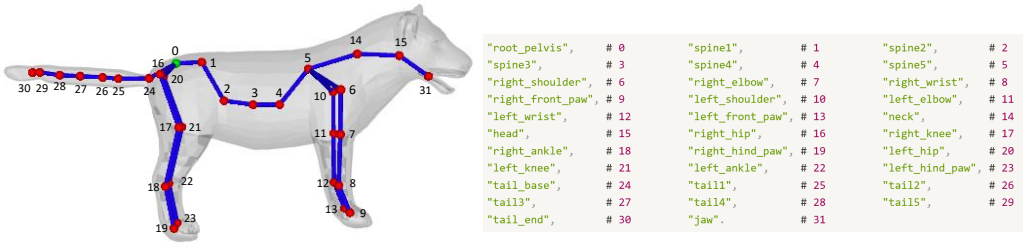

Figure 偏微分方程 8 : SMALR 数据 [106, 107] 的骨架结构与关节名称。

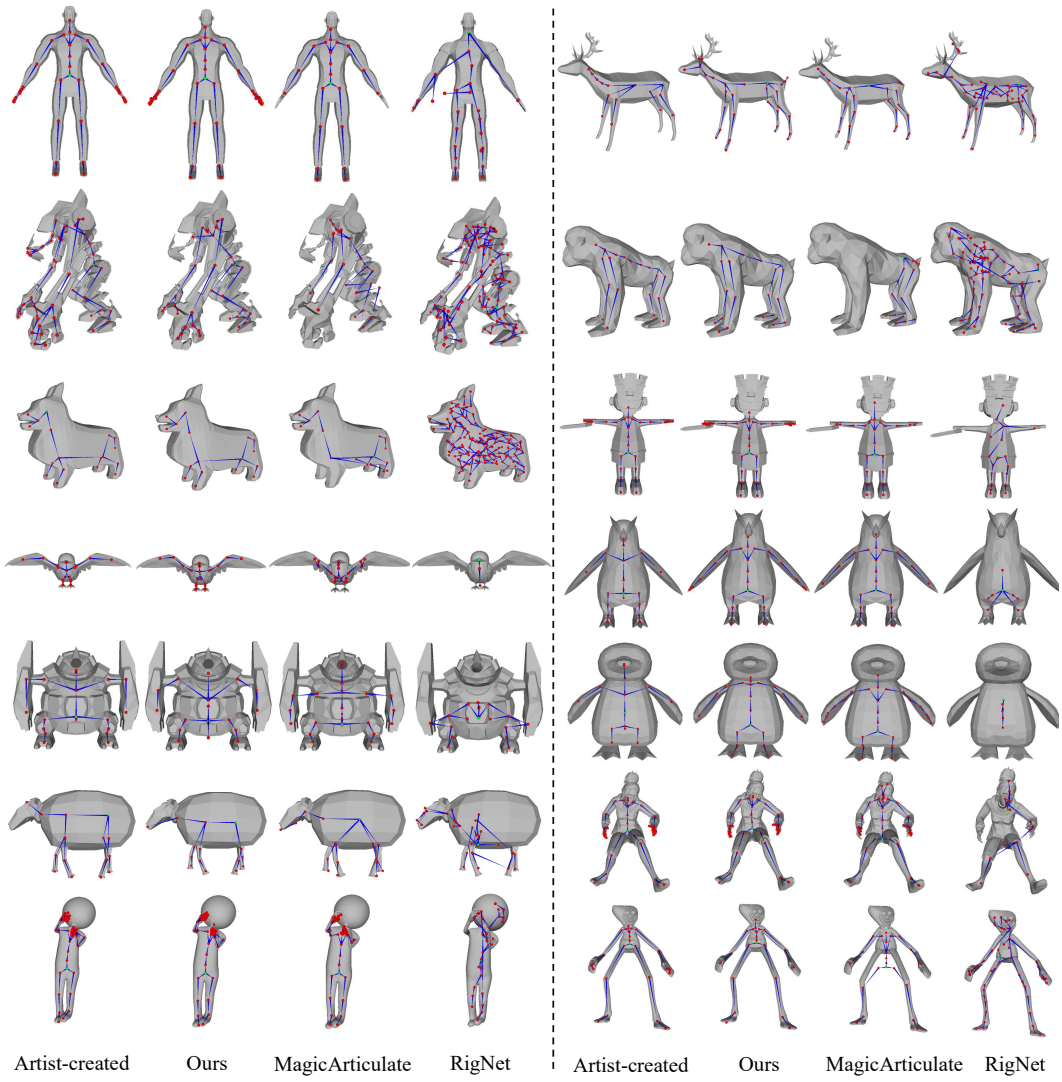

Figure 偏微分方程 9：对测试集上的骨架生成结果进行比较。从上到下：Articulation-XL2.0 中的六个例子，ModelsResource 中的四个，以及多样姿势子集中的四个。我们的方法生成了有效的骨架，甚至修正了艺术家创建的错误（例如，第 1 行缺失的鹿腿，第 4-5 行缺失的企鹅翅膀）。

C 附加实验结果

C.1 骨架生成的更多结果

More qualitative results on test sets. Figure 偏微分方程 9 对比了我们的方法 MagicArticulate [67] 和 RigNet [85] 对来自 Articulation-XL2.0 、ModelsResource 和多样化姿势子集的网格的骨架生成的额外定性比较。我们的方法产生了更有效的骨架，甚至可以纠正艺术家创建的错误，例如第一行中缺失的鹿腿，以及第 4 和第 5 行中缺失的企鹅手臂。

作为对主论文中消融研究的补充，**??** 比较了按层次顺序和空间顺序生成的骨架。使用空间顺序总会产生不连续的骨架，因为子关节在父关节之前生成，会创建无效的父级引用。

我们比较了在 Articulation-XL2.0 测试上的平均推断时间（见 Table S 6 ）。排除所有方法的数据预处理，我们的方法比 Pinocchio 2.6× 倍更快，比 RigNet 3.0× 倍更快，比 UniRig 1.9× 倍更快，以及比 MagicArticulate 1.6× 倍更快。

Table S 6 : 骨架生成的推理时间。

Method	Pinocchio	RigNet	UniRig	MagicArticulate	Ours
Inference time	3.9s	4.5s	2.9s	2.4s	1.5s

Table S 7 : 蒙皮权重预测的定量比较。我们将我们的方法与 GVB、RigNet 和 MagicArticulate 进行比较。对于平均 L1 范数误差（L1）和平均距离误差（avg Dist.），值越低越好。这里，* 表示在 Articulation-XL2.0 上使用多样姿势子集训练的模型。

Method	Articulation-XL2.0		ModelsResource		Diverse-pose	
	L1 ↓	avg Dist. ↓	L1 ↓	avg Dist. ↓	L1 ↓	avg Dist. ↓
GVB	0.745	0.0087	0.687	0.0067	0.786	0.0084
RigNet	0.729	0.0082	0.707	0.0078	0.746	0.0089
MagicArti.	0.451	0.0051	0.642	<u>0.0064</u>	0.479	0.0067
Ours	<u>0.335</u>	<u>0.0043</u>	<u>0.443</u>	0.0044	<u>0.405</u>	<u>0.0061</u>
Ours*	0.333	0.0042	0.442	0.0044	0.353	0.0053

C.2 蒙皮权重预测的更多结果

除了在主要论文中提出的用于评估蒙皮权重准确性的精度、召回率和 L1 范数之外，我们还通过形变误差分析进行了实际有效性的全面评估。这个补充指标量化了通过预测的蒙皮权重变形的顶点与通过真实权重变形的顶点之间的平均欧几里得距离，跨一个多样化的由 10 个随机生成的姿势组成的数据集。如 Table S 7 所示，所提出的方法在所有实验数据集上表现出了更好的性能。

除了在 **??** 中的消融结果，我们还对块深度进行消融。在 Figure 2 的绿色注意力块中，我们变化堆叠块的数量，我们称之为深度。当深度为 1（44.3M 参数）时，这对应于单个块。我们逐渐增加块深度，对应于更大的模型规模，并评估结果性能。结果如 Table S 8 所示：当深度设置为 2（87.7M 参数）时，Articulation-XL2.0 上的性能和多样化姿势子集明显改善，但在 ModelsResource 上略有下降。当深度为 3（130.9M 参数）时，Articulation-XL2.0 和多样化姿势子集的性能与深度 =2 相比结果相当，但在 ModelsResource 上仍然出现下降。我们将 ModelsResource 上的下降归因于 Articulation-XL2.0 和 ModelsResource 之间的方向分布差异：在使用高度变化的方向训练过 Articulation-XL2.0 后，更大的模型可能会偏向于多样化方向，导致在测试 ModelsResource 时性能下降，其中仅包含面向前方的方向。对于深度 =3，性能与深度 =2 相当而没有明显改善，这表明模型容量可能在当前数据集规模上已达到饱和。

More qualitative results. Figure 偏微分方程 10 进一步对比了我们的方法 MagicArticulate [67] 和 RigNet [85] 在来自 Articulation-XL2.0 、ModelsResource 和多样姿势子集的网格上的蒙皮权重预测的质性比较。每个例子将预测的权重图与其相对于艺术家绘制的参考的 L1 误差图进行配对，突出显示我们的方法在各种对象类别中的卓越精确度。

Inference time. 我们还比较了在 Articulation-XL2.0 测试上的蒙皮权重预测的平均推理时间（见 Table S 9 ）。除去数据预处理，我们的方法比 GVB 快 59× ，比 RigNet 快 1.75× ，比 MagicArticulate 快 45× 。

我们在 **??** 中展示了更多的动画结果。输入网格由 Tripo 2.0 [2] 和 Hunyuan3D 2.0 [73] 生成。尽管有很好对齐的参考视图，L4GM [59] 仍然会持续产生几何扭曲（用红色高亮显示），即便是在地面真实多视图渲染下也是如此。MotionDreamer [77] 生成了微妙的动画，并在像龟

Table S 8 : 关于蒙皮权重预测网络中注意力模块深度的消融研究。

Method	Articulation-XL2.0			ModelsResource			Diverse-pose		
	Prec. ↑	Rec. ↑	L1 ↓	Prec. ↑	Rec. ↑	L1 ↓	Prec. ↑	Rec. ↑	L1 ↓
depth = 1	87.6 %	74.0 %	0.335	<u>79.7 %</u>	81.6 %	0.443	83.6 %	72.2 %	0.405
depth = 2	<u>89.3 %</u>	<u>73.0 %</u>	0.316	79.8 %	<u>80.2 %</u>	<u>0.453</u>	85.8 %	<u>71.1 %</u>	<u>0.392</u>
depth = 3	89.4 %	72.5 %	<u>0.317</u>	79.6 %	79.7 %	0.461	<u>85.7 %</u>	70.8 %	0.391

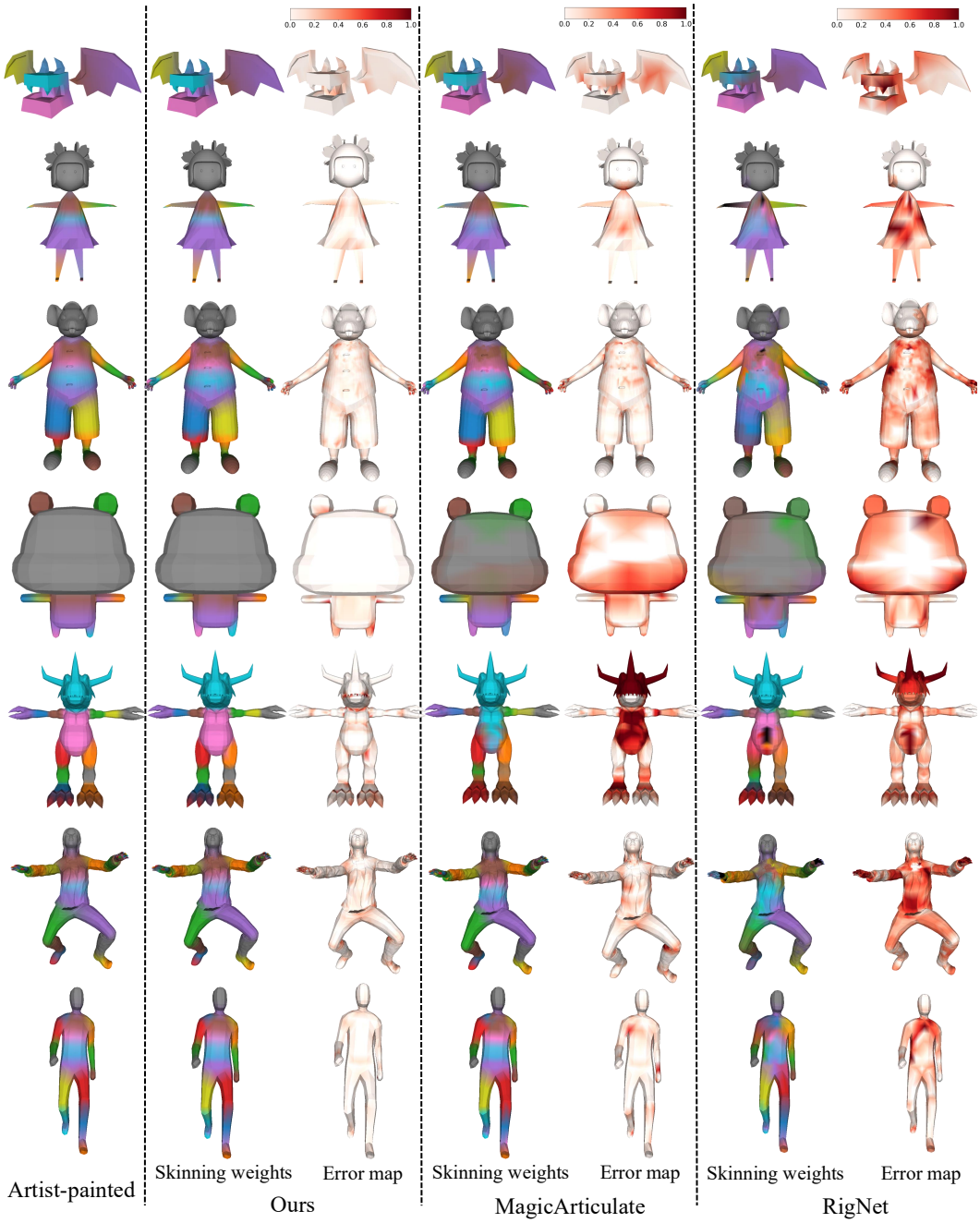

Figure 偏微分方程 10：皮肤权重预测结果的比较。从上到下：三个来自 Articulation-XL2.0，两个来自 ModelsResource，以及两个来自不同姿态子集的例子。每对显示预测权重的可视化及其 L1 误差图。我们的预测更加接近艺术家涂绘的参考。

Table S 9：蒙皮权重预测的推理时间。

Method	GVB	RigNet	MagicArticulate	Ours
Inference time	1.895s	0.056s	1.430s	0.032s

壳这样的刚性部件中引入了非预期的变形。相比之下，我们的方法使用我们生成的骨骼蒙皮产生准确、无伪影的动画。

我们进行了用户研究，研究对象为 21 名参与者，以评估三种方法的动画质量：L4GM [59] 、MotionDreamer [77] 、以及我们的方法。参与者针对三个评估标准比较了 8 个动画示例：（1）视频动画对齐度：哪个动画结果与输入视频的对齐度更好？（2）运动质量：哪个动画的运动更自然和真实？（3）三维几何保真度：哪个方法更好地保持了原始三维对象的几何形状而没有引入扭曲或伪影？结果如 Appendix C.2 所示。我们的方法在所有三个评估维度上均优于 L4GM 和 MotionDreamer。请注意，视频动画对齐度未针对 MotionDreamer 进行评估，因为它使用的是文本驱动的运动生成，而不是视频指导。

Table S 10 : 用户研究评估动画结果。

Method	Video-Animation Align.	Motion Quality	Geometry Preservation
MotionDreamer	-	0	0
L4GM	19.64 %	16.67 %	18.45 %
Ours	80.36 %	83.33 %	81.55 %

尽管其性能强大，我们的框架有两个主要限制。首先，它无法捕捉细微尺度的变形，例如流动的头发或飘动的布料，因为这些高度可变形的部分没有生成骨架。一个明显的例子是我们 项目页面 上的游泳海龟序列。虽然参考视频显示了海龟前肢的非常柔和的运动，我们的动画结果由于这些区域关节密度不足而显得不够流畅。这种限制源于我们的骨架生成在需要细微尺度变形的区域产生较少的关节。动画驱动的关节细化可以改善流畅度，但仍需未来工作。此外，动画阶段依赖于每个场景的优化，阻碍了实时部署。直接预测动画的端到端前馈模型可以消除这个瓶颈。

除了这些结构问题之外，一些实际因素也会影响动画质量。(1) 复杂的运动：快速的动作和大幅度的关节旋转带来了挑战。例如，在我们项目页面中的海马案例中，虽然我们捕捉到了整体尾部摆动的模式，但精确地对齐精细的运动仍然很困难。基于光流大小的自适应帧采样——在快速运动期间更密集地采样——可能可以解决这些挑战。(2) 视频生成质量：尽管现有的文本到视频模型（Kling AI [1] , JiMeng AI [34] ）可以高成功率生成复杂运动，但视频质量直接影响动画的逼真度。运动模糊或时间不一致性会降低关节/顶点跟踪的准确性，并使优化更具挑战性。我们通过生成多个视频候选并基于视觉清晰度和运动一致性选择质量最高的一个来缓解这一问题。虽然这种方法有助于减少劣质视频的影响，但视频生成质量对于高度复杂的运动场景仍然是一个限制。(3) 视点和遮挡问题：不理想的摄像机角度可能导致深度模糊和跟踪失败。虽然我们可以使用输入的 3D 网格选择最大化关节可见性的最佳视角，但当关键关节在整个序列中保持遮挡时，单视图优化本质上会遇到困难。多视图先验可能会对遮挡区域提供更好的几何理解。

D 更广泛的影响

除了技术上的突破，这项工作通过消除对专业知识的需求，赋予曾被动画工具排除在外的多样化创作者力量，具有重要的社会意义。随着数字环境日益影响我们的工作、学习和沟通方式，扩大对动画内容创作的访问不仅在技术上具有价值，而且在社会上是必不可少的。然而，动画技术的普及也引发了关于可能在创建欺骗性内容中的误用以及对传统动画行业就业影响的担忧。我们的最终愿景是将 3D 动画从一种专属的专业技能转变为一种直观的创作媒介，让每个人都能使用，同时鼓励负责任的使用。

References

[1] K. AI. Kling ai, 2025. URL https://klingai.com/.

[2] T. AI. Tripo 3d, 2023. URL https://www.tripo3d.ai/.

[3] O. K.-C. Au, C.-L. Tai, H.-K. Chu, D. Cohen-Or, and T.-Y. Lee. Skeleton extraction by mesh contraction. *ACM transactions on graphics (TOG)*, 27(3):1–10, 2008.

[4] S. Bahmani, X. Liu, W. Yifan, I. Skorokhodov, V. Rong, Z. Liu, X. Liu, J. J. Park, S. Tulyakov, G. Wetzstein, A. Tagliasacchi, and D. B. Lindell. Tc4d: Trajectory-conditioned text-to-4d generation. *arXiv*, 2024.

[5] S. Bahmani, I. Skorokhodov, V. Rong, G. Wetzstein, L. Guibas, P. Wonka, S. Tulyakov, J. J. Park, A. Tagliasacchi, and D. B. Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7996–8006, 2024.

[6] I. Baran and J. Popović. Automatic rigging and animation of 3d characters. *ACM Transactions on graphics (TOG)*, 26(3):72–es, 2007.

[7] J. Cao, A. Tagliasacchi, M. Olson, H. Zhang, and Z. Su. Point cloud skeletons via laplacian based contraction. In *2010 Shape Modeling International Conference*, pages 187–197. IEEE, 2010.

[8] C. Chen, S. Huang, X. Chen, G. Chen, X. Han, K. Zhang, and M. Gong. Ct4d: Consistent text-to-4d generation with animatable meshes. *arXiv preprint arXiv:2408.08342*, 2024.

[9] J. Chen, B. Zhang, X. Tang, and P. Wonka. V2m4: 4d mesh animation reconstruction from a single monocular video. *arXiv preprint arXiv:2503.09631*, 2025.

[10] S. Chen, H. Guo, S. Zhu, F. Zhang, Z. Huang, J. Feng, and B. Kang. Video depth anything: Consistent depth estimation for super-long videos. *arXiv:2501.12375*, 2025.

[11] Y. Chen, T. He, D. Huang, W. Ye, S. Chen, J. Tang, X. Chen, Z. Cai, L. Yang, G. Yu, et al. Meshanything: Artist-created mesh generation with autoregressive transformers. *arXiv preprint arXiv:2406.10163*, 2024.

[12] Y. Chen, Y. Wang, Y. Luo, Z. Wang, Z. Chen, J. Zhu, C. Zhang, and G. Lin. Meshanything v2: Artist-created mesh generation with adjacent mesh tokenization. *arXiv preprint arXiv:2408.02555*, 2024.

[13] Z. Chu, F. Xiong, M. Liu, J. Zhang, M. Shao, Z. Sun, D. Wang, and M. Xu. Humanrig: Learning automatic rigging for humanoid character in a large scale dataset, 2024. URL https://arxiv.org/abs/2412.02317.

[14] E. De Aguiar, C. Theobalt, S. Thrun, and H.-P. Seidel. Automatic conversion of mesh animations into skeleton-based animations. In *Computer Graphics Forum*, volume 27, pages 389–397. Wiley Online Library, 2008.

[15] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023.

[16] M. Deitke, R. Liu, M. Wallingford, H. Ngo, O. Michel, A. Kusupati, A. Fan, C. Laforte, V. Voleti, S. Y. Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024.

[17] Y. Deng, Y. Zhang, C. Geng, S. Wu, and J. Wu. Anymate: A dataset and baselines for learning 3d object rigging. In *SIGGRAPH*, 2025.

[18] O. Dionne and M. de Lasa. Geodesic voxel binding for production character meshes. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 173–180, 2013.

[19] A. Dodik, V. Sitzmann, J. Solomon, and O. Stein. Robust biharmonic skinning using geometric fields. *arXiv preprint arXiv:2406.00238*, 2024.

[20] Z. Fu, J. Wei, W. Shen, C. Song, X. Yang, F. Liu, X. Yang, and G. Lin. Sync4d: Video guided controllable dynamics for physics-based 4d generation. *arXiv preprint arXiv:2405.16849*, 2024.

[21] Q. Gao, Q. Xu, Z. Cao, B. Mildenhall, W. Ma, L. Chen, D. Tang, and U. Neumann. Gaussianflow: Splatting gaussian dynamics for 4d content creation. 2024.

[22] I. Gat, S. Raab, G. Tevet, Y. Reshef, A. H. Bermano, and D. Cohen-Or. Anytop: Character animation diffusion with any topology. *arXiv preprint arXiv:2502.17327*, 2025.

[23] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5152–5161, 2022.

[24] Z. Guo, J. Xiang, K. Ma, W. Zhou, H. Li, and R. Zhang. Make-it-animatable: An efficient framework for authoring animation-ready 3d characters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[25] H. Huang, S. Wu, D. Cohen-Or, M. Gong, H. Zhang, G. Li, and B. Chen. L1-medial skeleton of point cloud. *ACM Trans. Graph.*, 32(4):65–1, 2013.

[26] A. Inc. Mixamo. URL https://www.mixamo.com/.

[27] A. Inc. Autodesk maya, 2024. URL https://www.autodesk.com/products/maya/overview. Version 2024.

[28] A. Jacobson, I. Baran, J. Popovic, and O. Sorkine. Bounded biharmonic weights for real-time deformation. *ACM Trans. Graph.*, 30(4):78, 2011.

[29] A. Jacobson, D. Panozzo, et al. libigl: A simple C++ geometry processing library, 2018. https://libigl.github.io/.

[30] D. L. James and C. D. Twigg. Skinning mesh animations. *ACM Transactions on Graphics (TOG)*, 24(3):399–407, 2005.

[31] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, and T. Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023.

[32] Y. Jiang, L. Zhang, J. Gao, W. Hu, and Y. Yao. Consistent4d: Consistent 360 {\deg} dynamic object generation from monocular video. *arXiv preprint arXiv:2311.02848*, 2023.

[33] Y. Jiang, C. Yu, C. Cao, F. Wang, W. Hu, and J. Gao. Animate3d: Animating any 3d model with multi-view video diffusion. *arXiv preprint arXiv:2407.11398*, 2024.

[34] JiMeng AI. Jimeng ai, 2025. URL https://jimeng.jianying.com.

[35] N. Karaev, I. Makarov, J. Wang, N. Neverova, A. Vedaldi, and C. Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. *arXiv preprint arXiv:2410.11831*, 2024.

[36] J. P. Lewis, M. Cordner, and N. Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 811–818. 2023.

[37] P. Li, K. Aberman, R. Hanocka, L. Liu, O. Sorkine-Hornung, and B. Chen. Learning skeletal articulations with neural blend shapes. *ACM Transactions on Graphics (TOG)*, 40(4):1–15, 2021.

[38] X. Li, Q. Ma, T.-Y. Lin, Y. Chen, C. Jiang, M.-Y. Liu, and D. Xiang. Articulated kinematics distillation from video diffusion models. *arXiv preprint arXiv:2504.01204*, 2025. URL https://arxiv.org/abs/2504.01204.

[39] Z. Li, Y. Chen, and P. Liu. Dreammesh4d: Video-to-4d generation with sparse-controlled gaussian-mesh hybrid representation. *Advances in Neural Information Processing Systems*, 37:21377–21400, 2024.

[40] Z. Li, D. Litvak, R. Li, Y. Zhang, T. Jakab, C. Rupprecht, S. Wu, A. Vedaldi, and J. Wu. Learning the 3d fauna of the web. In *CVPR*, 2024.

[41] H. Liang, Y. Yin, D. Xu, H. Liang, Z. Wang, K. N. Plataniotis, Y. Zhao, and Y. Wei. Diffusion4d: Fast spatial-temporal consistent 4d generation via video diffusion models. *arXiv preprint arXiv:2405.16645*, 2024.

[42] C. Lin, C. Li, Y. Liu, N. Chen, Y.-K. Choi, and W. Wang. Point2skeleton: Learning skeletal representations from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4277–4286, 2021.

[43] J. Lin, Z. Wang, Y. Hou, Y. Tang, and M. Jiang. Phy124: Fast physics-driven 4d content generation from a single image. *arXiv preprint arXiv:2409.07179*, 2024.

[44] H. Ling, S. W. Kim, A. Torralba, S. Fidler, and K. Kreis. Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8576–8588, 2024.

[45] I. Liu, Z. Xu, W. Yifan, H. Tan, Z. Xu, X. Wang, H. Su, and Z. Shi. Riganything: Template-free autoregressive rigging for diverse 3d assets. *arXiv preprint arXiv:2502.09615*, 2025.

[46] L. Liu, Y. Zheng, D. Tang, Y. Yuan, C. Fan, and K. Zhou. Neuroskinning: Automatic skin binding for production characters with deep graph networks. *ACM Transactions on Graphics (ToG)*, 38(4):1–12, 2019.

[47] M. Liu, M. A. Uy, D. Xiang, H. Su, S. Fidler, N. Sharp, and J. Gao. Partfield: Learning 3d feature fields for part segmentation and beyond. *arXiv preprint arXiv:2504.11451*, 2025.

[48] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015.

[49] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. 1998.

[50] Q. Miao, J. Quan, K. Li, and Y. Luo. Pla4d: Pixel-level alignments for text-to-4d gaussian splatting. *arXiv preprint arXiv:2405.19957*, 2024.

[51] Q. Miao, K. Li, J. Quan, Z. Min, S. Ma, Y. Xu, Y. Yang, and Y. Luo. Advances in 4d generation: A survey. *arXiv preprint arXiv:2503.14501*, 2025.

[52] M. B. S. Millán, A. Dai, and M. Nießner. Animating the uncaptured: Humanoid mesh animation with video diffusion models. *arXiv preprint arXiv:2503.15996*, 2025.

[53] H. Morimitsu, X. Zhu, R. M. Cesar-Jr., X. Ji, and X.-C. Yin. DPFlow: Adaptive optical flow estimation with a dual-pyramid framework. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[54] A. Mosella-Montoro and J. Ruiz-Hidalgo. Skinningnet: Two-stream graph convolutional neural network for skinning prediction of synthetic characters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18593–18602, 2022.

[55] X. Pan, J. Huang, J. Mai, H. Wang, H. Li, T. Su, W. Wang, and X. Jin. Heterskinnet: A heterogeneous network for skin weights prediction. *Proc. ACM Comput. Graph. Interact. Tech.*, 4(1), Apr. 2021. doi: 10.1145/3451262. URL https://doi.org/10.1145/3451262.

[56] R. Parent. *Computer animation: algorithms and techniques*. Newnes, 2012.

[57] 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.

[58] J. Ren, L. Pan, J. Tang, C. Zhang, A. Cao, G. Zeng, and Z. Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023.

[59] J. Ren, C. Xie, A. Mirzaei, K. Kreis, Z. Liu, A. Torralba, S. Fidler, S. W. Kim, H. Ling, et al. L4gm: Large 4d gaussian reconstruction model. *Advances in Neural Information Processing Systems*, 37:56828–56858, 2024.

[60] W. Shen, W. Yin, H. Wang, C. Wei, Z. Cai, L. Yang, and G. Lin. Hmr-adapter: A lightweight adapter with dual-path cross augmentation for expressive human mesh recovery. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6093–6102, 2024.

[61] W. Shen, W. Yin, X. Yang, C. Chen, C. Song, Z. Cai, L. Yang, H. Wang, and G. Lin. Adhmr: Aligning diffusion-based human mesh recovery via direct preference optimization. *arXiv preprint arXiv:2505.10250*, 2025.

[62] Y. Siddiqui, A. Alliegro, A. Artemov, T. Tommasi, D. Sirigatti, V. Rosov, A. Dai, and M. 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.

[63] C. Song, J. Wei, R. Li, F. Liu, and G. Lin. 3d pose transfer with correspondence learning and mesh refinement. *Advances in Neural Information Processing Systems*, 34:3108–3120, 2021.

[64] C. Song, J. Wei, R. Li, F. Liu, and G. Lin. Unsupervised 3d pose transfer with cross consistency and dual reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10488–10499, 2023.

[65] C. Song, J. Wei, T. Chen, Y. Chen, C.-S. Foo, F. Liu, and G. Lin. Moda: Modeling deformable 3d objects from casual videos. *International Journal of Computer Vision*, pages 1–20, 2024.

[66] C. Song, J. Wei, C. S. Foo, G. Lin, and F. Liu. Reacto: Reconstructing articulated objects from a single video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5384–5395, 2024.

[67] C. Song, J. Zhang, X. Li, F. Yang, Y. Chen, Z. Xu, J. H. Liew, X. Guo, F. Liu, J. Feng, and G. Lin. Magicarticulate: Make your 3d models articulation-ready. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[68] K. Sun, D. Litvak, Y. Zhang, H. Li, J. Wu, and S. Wu. Ponymation: Learning articulated 3d animal motions from unlabeled online videos. In *ECCV*, 2024.

[69] M. Sun, J. Chen, J. Dong, Y. Chen, X. Jiang, S. Mao, P. Jiang, J. Wang, B. Dai, and R. Huang. Drive: Diffusion-based rigging empowers generation of versatile and expressive characters. *arXiv preprint arXiv:2411.17423*, 2024.

[70] Q. Sun, Z. Guo, Z. Wan, J. N. Yan, S. Yin, W. Zhou, J. Liao, and H. Li. Eg4d: Explicit generation of 4d object without score distillation. *arXiv preprint arXiv:2405.18132*, 2024.

[71] A. Tagliasacchi, I. Alhashim, M. Olson, and H. Zhang. Mean curvature skeletons. In *Computer Graphics Forum*, volume 31, pages 1735–1744. Wiley Online Library, 2012.

[72] J. Tang, Z. Li, Z. Hao, X. Liu, G. Zeng, M.-Y. Liu, and Q. Zhang. Edgerunner: Autoregressive auto-encoder for artistic mesh generation. *arXiv preprint arXiv:2409.18114*, 2024.

[73] T. H. Team. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation, 2025.

[74] G. Tevet, B. Gordon, A. Hertz, A. H. Bermano, and D. Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer, 2022.

[75] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, and A. H. Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

[76] The Models-Resource. The models-resource, 2019. URL https://www.models-resource.com/.

[77] L. Uzolas, E. Eisemann, and P. Kellnhofer. Motiondreamer: Exploring semantic video diffusion features for zero-shot 3d mesh animation. In *International Conference on 3D Vision 2025*, 2025.

[78] P.-S. Wang, Y. Liu, and X. Tong. Dual octree graph networks for learning adaptive volumetric shape representations. *ACM Transactions on Graphics (TOG)*, 41(4):1–15, 2022.

[79] X. Wang, Y. Wang, J. Ye, F. Sun, Z. Wang, L. Wang, P. Liu, K. Sun, X. Wang, W. Xie, et al. Animatabledreamer: Text-guided non-rigid 3d model generation and reconstruction with canonical score distillation. In *European Conference on Computer Vision*, pages 321–339. Springer, 2024.

[80] S. Wu, R. Li, T. Jakab, C. Rupprecht, and A. Vedaldi. MagicPony: Learning articulated 3d animals in the wild. In *CVPR*, 2023.

[81] Z. Wu, C. Yu, Y. Jiang, C. Cao, F. Wang, and X. Bai. Sc4d: Sparse-controlled video-to-4d generation and motion transfer. In *European Conference on Computer Vision*, pages 361–379. Springer, 2024.

[82] Y. Xie, C.-H. Yao, V. Voleti, H. Jiang, and V. Jampani. Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. *arXiv preprint arXiv:2407.17470*, 2024.

[83] J. L. Z. W. D. Xu and S. J. Y. G. M. Jiang. Phys4dgen: Physics-compliant 4d generation with multi-material composition perception.

[84] Z. Xu, Y. Zhou, E. Kalogerakis, and K. Singh. Predicting animation skeletons for 3d articulated models via volumetric nets. In *2019 international conference on 3D vision (3DV)*, pages 298–307. IEEE, 2019.

[85] Z. Xu, Y. Zhou, E. Kalogerakis, C. Landreth, and K. Singh. Rignet: Neural rigging for articulated characters. *arXiv preprint arXiv:2005.00559*, 2020.

[86] Z. Xu, Y. Zhou, L. Yi, and E. Kalogerakis. Morig: Motion-aware rigging of character meshes from point clouds. In *SIGGRAPH Asia 2022 conference papers*, pages 1–9, 2022.

[87] G. Yang, M. Vo, N. Neverova, D. Ramanan, A. Vedaldi, and H. Joo. Banmo: Building animatable 3d neural models from many casual videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2863–2873, 2022.

[88] C.-H. Yao, Y. Xie, V. Voleti, H. Jiang, and V. Jampani. Sv4d 2.0: Enhancing spatio-temporal consistency in multi-view video diffusion for high-quality 4d generation. *arXiv preprint arXiv:2503.16396*, 2025.

[89] Y. Yao, Z. Deng, and J. Hou. Riggs: Rigging of 3d gaussians for modeling articulated objects in videos. In *CVPR*, 2025.

[90] Y. Yin, D. Xu, Z. Wang, Y. Zhao, and Y. Wei. 4dgen: Grounded 4d content generation with spatial-temporal consistency. *arXiv preprint arXiv:2312.17225*, 2023.

[91] Q. Yu, J. He, X. Deng, X. Shen, and L.-C. Chen. Randomized autoregressive visual generation. *arXiv preprint arXiv:2411.00776*, 2024.

[92] Y.-J. Yuan, L. Kobbelt, J. Liu, Y. Zhang, P. Wan, Y.-K. Lai, and L. Gao. 4dynamic: Text-to-4d generation with hybrid priors. *arXiv preprint arXiv:2407.12684*, 2024.

[93] K. Yun, S. Hong, C. Kim, and J. Noh. Anymole: Any character motion in-betweening leveraging video diffusion models. *arXiv preprint arXiv:2503.08417*, 2025.

[94] B. Zeng, L. Yang, S. Li, J. Liu, Z. Zhang, J. Tian, K. Zhu, Y. Guo, F.-Y. Wang, M. Xu, et al. Trans4d: Realistic geometry-aware transition for compositional text-to-4d synthesis. *arXiv preprint arXiv:2410.07155*, 2024.

[95] Y. Zeng, Y. Jiang, S. Zhu, Y. Lu, Y. Lin, H. Zhu, W. Hu, X. Cao, and Y. Yao. Stag4d: Spatial-temporal anchored generative 4d gaussians. In *European Conference on Computer Vision*, pages 163–179. Springer, 2024.

[96] B. Zhang and P. Wonka. Functional diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4723–4732, 2024.

[97] H. Zhang, D. Chang, F. Li, M. Soleymani, and N. Ahuja. Magicpose4d: Crafting articulated models with appearance and motion control. *arXiv preprint arXiv:2405.14017*, 2024.

[98] H. Zhang, X. Chen, Y. Wang, X. Liu, Y. Wang, and Y. Qiao. 4diffusion: Multi-view video diffusion model for 4d generation. *Advances in Neural Information Processing Systems*, 37: 15272–15295, 2024.

[99] J.-P. Zhang, C.-F. Pu, M.-H. Guo, Y.-P. Cao, and S.-M. Hu. One model to rig them all: Diverse skeleton rigging with unirig. *arXiv preprint arXiv:2504.12451*, 2025.

[100] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence*, 46(6):4115–4128, 2024.

[101] M. Zhang, D. Jin, C. Gu, F. Hong, Z. Cai, J. Huang, C. Zhang, X. Guo, L. Yang, Y. He, et al. Large motion model for unified multi-modal motion generation. In *European Conference on Computer Vision*, pages 397–421. Springer, 2024.

[102] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[103] Y. Zhao, Z. Yan, E. Xie, L. Hong, Z. Li, and G. H. Lee. Animate124: Animating one image to 4d dynamic scene. *arXiv preprint arXiv:2311.14603*, 2023.

[104] Z. Zhao, W. Liu, X. Chen, X. Zeng, R. Wang, P. Cheng, B. Fu, T. Chen, G. Yu, and S. Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. *Advances in Neural Information Processing Systems*, 36, 2024.

[105] H. Zhu, T. He, X. Yu, J. Guo, Z. Chen, and J. Bian. Ar4d: Autoregressive 4d generation from monocular videos. *arXiv preprint arXiv:2501.01722*, 2025.

[106] S. Zuffi, A. Kanazawa, D. Jacobs, and M. J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[107] S. Zuffi, A. Kanazawa, and M. J. Black. Lions and tigers and bears: Capturing non-rigid, 3D, articulated shape from images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2018.