arXiv:2601.06378v1 [cs.GR] 10 Jan 2026

RigMo：统一绑定与运动学习，实现生成式动画

Hao Zhang[1,2] Jiahao Luo[1,3] Bohui Wan[2] Yizhou Zhao[1,4] Zongrui Li[5]Michael Vasilkovsky[1] Chaoyang Wang[1] Jian Wang[1] Narendra Ahuja[2] Bing Zhou[1]Snap Inc. [2]伊利诺伊大学厄巴纳-香槟分校 [3]加利福尼亚大学圣克鲁兹分校[4]卡内基梅隆大学 [5]南洋理工大学

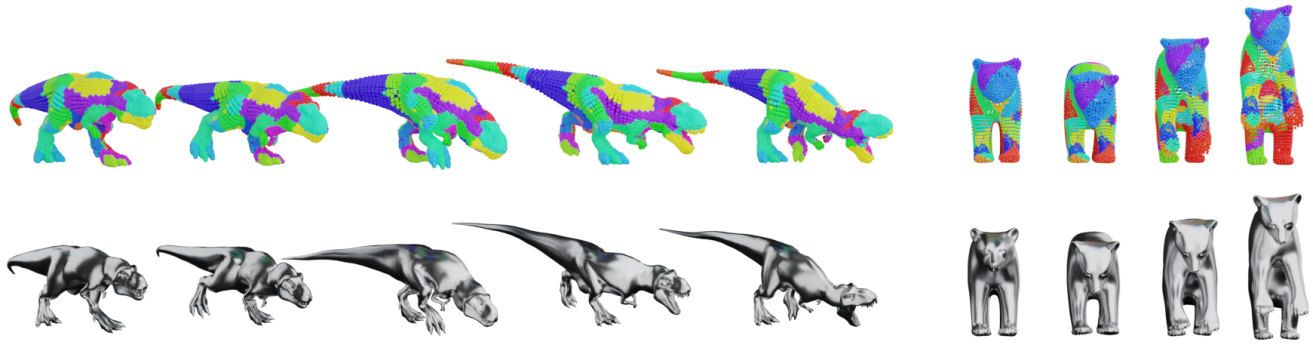


图 1。**RigMo** 通过理解网格序列的底层结构，联合学习绑定与运动。不同于为每个序列拟合绑定的基于优化的方法，RigMo是一个前馈框架，可直接从输入网格中推断高斯骨骼、蒙皮权重和运动参数，从而实现统一的 4D动画生成。颜色展示了高斯骨骼对网格表面顶点的影响（蒙皮权重）；由于高斯骨骼是为可视化随机分配的，不同骨骼可能呈现相似的颜色。

摘要

尽管*4D*生成取得了显著进展，绑定与运动——动画的核心结构和动态组成部分——通常仍被视为相互独立的问题。现有流程依赖真实骨架和蒙皮权重来进行运动生成，并将自动绑定视为独立过程，从而削弱了可扩展性与可解释性。我们提出了***RigMo***，这是一种统一的生成框架，能够直接从原始网格序列中联合学习绑定与运动，无需任何人工提供的绑定标注。*RigMo*将逐顶点变形编码到两个紧凑的潜在空间中：绑定潜变量解码为显式的高斯骨骼和蒙皮权重，运动潜变量则生成随时间变化的*SE(3)*变换。二者共同定义了具有显式结构和连贯运动的可动画化网格，从而能够对可变形物体进行前馈式骨架与运动推断。除了统一的绑定—运动发现之外，我们还引入了一个运行于*RigMo*潜在空间中的*Motion-DiT*模型，并证明这些结构感知潜变量能够自然地支持下游运动生成任务。在*DeformingThings4D*、*Objaverse-XL*和*TrueBones*上的实验表明，*RigMo*能够学习平滑、可解释且符合物理规律的绑定，并且与现有自动绑定和变形基线相比，在重建效果和类别级泛化能力方面表现更优，从而表明*RigMo*能够学习平滑、可解释且符合物理规律的绑定，同时相较于现有自动绑定和变形基线，在重建效果和类别级泛化能力方面表现更优。*RigMo*确立了统一、结构感知且可扩展的动态*3D*建模新范式。项目页面：*https://RigMo-Page.github.io*

1. 引言

动画本质上将结构与运动联系在一起。结构通常由绑定表示，用于定义物体如何发生形变；而运动则描述该结构如何随时间演化。然而，尽管二者具有内在的相互依赖性，现有的大多数流程仍将这些组成部分割裂开来处理。

自动绑定系统 [20, 29] 依赖艺术家设计的骨架和蒙皮权重，训练模型模仿人类的启发式方法，而非理解为何某些绑定能产生合理的形变。因此，这些方法高度依赖人工标注，限制了它们在不同数据集和物体类别中的可扩展性与一致性。第二类运动生成方法假设

已提供真实骨架。人体运动模型 [1, 11, 16, 22, 24, 25, 33, 34, 41, 44, 55, 57] 和以动物或物体为中心的流程 [26, 58]（例如 Any-Top） [9] 完全在姿态空间中运行，在预定义的运动学结构上预测关节旋转或 SE(3) 变换。这些方法无法推断骨架结构，无法处理任意几何形状；当假设的骨架不匹配或不可用时，也会失效。第三类工作则完全摒弃绑定。现代顶点空间运动生成器和 4D 重建方法 [2, 3, 8, 13, 28, 38,46, 53]，（包括 AnimateAnyMesh） [38]和 GVFDiffusion） [46], 逐帧预测形变，不进行任何结构抽象。这些模型虽然灵活，却难以控制、难以解释，也无法生成可复用的可动画资产——而这正是绑定的核心目的。综上，这些范式揭示了一个根本性缺口：目前尚无统一框架能够直接从原始网格序列中联合学习骨架结构与运动动力学，同时无需预定义骨架或针对每个序列进行优化。与此同时，依赖人工标注骨架成本高且不一致，因此，完全由数据驱动的替代方案既必要又实用。这促使我们开发一种模型，使其能够从形变中发现结构，并生成符合所学结构的运动。

我们提出 RigMo，这是一种通过统一架构联合学习绑定与运动、从而弥补这一空白的生成模型。RigMo 将动态网格序列编码到两个互补的潜在空间中，分别捕捉空间关节结构和时间演化。解码器重建一组显式的高斯骨骼 [42], ，并由此推导出蒙皮权重，同时预测每一帧的 SE(3) 变换，共同生成一种可完全驱动动画的表示。这种高斯骨骼表述提供了一种紧凑且可扩展的抽象，其复杂度仅取决于骨骼数量，而非网格分辨率，并且本质上不受分辨率影响。此外，该模型以自监督方式训练，无需人工标注，因此能够利用 DeformingThings4D [18] 等未标注数据集进行学习。

与经典的刚性骨骼平滑蒙皮分解（SSDR）不同，后者需要针对每个序列分别求解优化问题；RigMo 执行前馈式骨架与运动推断，并能泛化到未见过的类别和运动风格。由此得到的表示是显式、可解释且可直接操控的，支持运动生成、运动插值和可控编辑等下游应用。这也自然引出了我们的第二项贡献。

除了统一发现绑定与运动之外，我们还提出了 Motion-DiT，这是一种在 RigMo 的运动潜在空间中运行的扩散 Transformer。Motion-DiT 不直接生成原始顶点坐标，而是在结构感知潜在域中生成时间轨迹，从而为实现可控、高质量的动作合成提供了原则性途径——并展示了 RigMo 学得的潜在表示对下游动作生成任务的实用价值。

在 DeformingThings4D、Objaverse-XL [7]，和 TrueBones 上，RigMo 直接从原始形变数据中恢复出物理上合理的骨架、连贯的蒙皮结构和逼真的运动，从而建立了一个统一、具备结构感知能力且可控的 4D生成动画框架。

2. 相关工作

4D生成模型。 近期研究提出了多种用于 4D生成的生成模型 [5, 6, 8, 12, 19,27, 35, 37, 39, 43, 45, 47, 48, 56]。一种常见框架是使用精心设计的 3D 或 4D 变分自编码器，将几何和运动嵌入潜在空间，再由生成网络（即，扩散模型 [10]）根据文本或视频条件对这些潜变量进行解码。尽管这类方法能够生成视觉效果出色的结果，但生成的运动往往缺乏可控性：它们生成的动画没有显式绑定及相应的运动参数，这极大限制了其在实际应用中的适用性。在本工作中，我们通过将可变形网格编码为统一的时空潜在表示，来解决这一问题；这些表示还可进一步解码为高度可控的骨架结构及其关联运动。

自动绑定。 自动绑定旨在通过预测骨架和逐顶点蒙皮权重，将静态网格转化为可动画化资产。基于模板的方法 [4, 15,23] 会附加预定义的骨骼结构，并优化其姿势和权重，但在处理人形以外的形状时泛化能力较差。RigNet [40] 等无模板方法放宽了对关节拓扑的假设，但可能生成过于密集的骨架，或依赖手工设计的几何线索。近期，自回归模型 [21, 30, 31, 52]将骨架生成视为序列预测问题，并借助 Articulation-XL [31]等大型数据集，学习类别无关的关节先验。尽管取得了进展，这些方法仍依赖大量由艺术家标注的绑定，而此类数据仍然难以获取且成本高昂。

基于优化的逆向蒙皮。 逆向蒙皮 [32, 36, 49–51]研究的是自动绑定的互补任务：给定一系列关节式网格形变，目标是恢复底层绑定参数，包括骨骼变换、蒙皮权重或弹性属性。基于线性混合蒙皮（LBS）的经典解决方案计算效率高，但将形变限制为变换的线性混合，从而限制了真实感。近期基于物理原理的公式 [51]联合优化骨骼运动和材料属性，从而能够实现非线性体积行为。然而，这些流程需要针对每个案例进行 非线性优化，

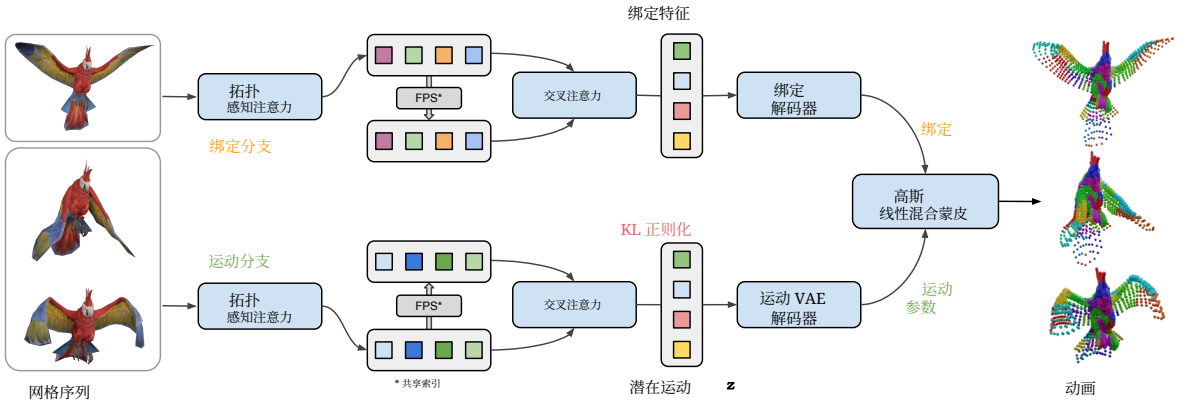


图 2. **RigMo-VAE** 框架概览。给定形变网格序列中的顶点时序轨迹，RigMo 采用双路径编码器，将静态几何（绑定分支）与动态运动（运动分支）解耦，学习同时捕捉空间结构和时序动态的紧凑潜在表示。解码器将这些潜在特征映射为具有物理可解释性的绑定组件：定义测地感知蒙皮权重的高斯骨骼描述符，以及用于局部变换和根变换的变分运动参数。不同颜色表示所学习高斯骨骼的影响区域，展示了无需人工绑定监督即可实现的、具有语义意义的网格形变分解。

因此速度缓慢且难以扩展。此外，恢复出的绑定与各自的序列紧密关联，无法泛化到不同主体或运动风格。相比之下，我们的方法在单次前馈过程中同时预测运动轨迹和隐式骨架参数，并从所有训练序列中联合学习，从而获得一种共享且可泛化的关节表示，能够稳健地迁移到各种物体和运动。

3. 方法

RigMo 是一种新颖的框架，通过学习由高斯骨骼组成的绑定表示来建模可变形表面运动（参见第 3.1.3 节中的高斯骨骼定义）$\mathbf{G}$及其关联的运动参数$\{\mathbf{q}_{\text{local}}, \mathbf{t}_{\text{local}}, \mathbf{q}_{\text{root}}, \mathbf{t}_{\text{root}}\}$。其中 $\mathbf{q}$ 表示旋转四元数，$\mathbf{t}$表示平移向量；二者通过可微蒙皮共同控制网格变形。不同于需要针对每个案例进行拟合的 SSDR 方法，RigMo 是一种前馈框架，能够直接从输入序列中推断绑定结构和运动。

我们的方法由两个主要组件构成：**RigMo VAE**（第 3.1 节），用于学习绑定表示及相应的运动参数；以及 **MotionDiT** 模块（第 3.2 节），该模块在 RigMo VAE 学得的潜在空间中运行，并将骨架分支输出作为条件信号，对运动潜变量进行扩散，从而生成可控动画。

3.1. RigMo 变分自编码器

给定一个网格序列$\mathbf{V} \in \mathbb{R}^{B\times T\times N\times 3}$，其批量大小为$B$、帧数为 T，且每个网格有 N 个顶点，RigMo 从能够同时编码空间结构和时序动态的潜在表示中重建可变形运动 ^<style id='7'>V </style>{v9}。网络架构由以下部分组成：（1）一个拓扑感知编码器，用于聚合几何信息和运动线索；（2）一个解码器，用于预测高斯骨骼及其变换；（3）一个基于高斯的线性混合蒙皮（LBS）模块，在学习得到的绑定参数下合成形变。

3.1.1. 拓扑感知编码器

RigMo 的编码器通过双路径架构联合捕获静态几何与动态运动，旨在将规范形状信息与时间形变线索解耦。这确保绑定预测反映的是稳定结构，而非特定运动实例，从而实现可泛化、一致且可解释的骨骼—顶点关系。

绑定分支。 绑定分支处理规范网格几何（第一帧），以建立骨骼—顶点对应关系并预测高斯骨骼参数。我们通过拓扑感知注意力将输入顶点 $\mathbf{V}_0 \in \mathbb{R}^{N\times 3}$ 编码为逐顶点嵌入 $\mathbf{V}_0^{\text{emb}}$：

$$\mathbf{h}_\ell = \text{Attn}(\text{LN}(\mathbf{h}_{\ell-1}), \mathcal{N}) + \mathbf{h}_{\ell-1}. \tag{1}$$

最远点采样选取 K 个骨骼令牌$\mathbf{B}^{\text{emb}}$ 及坐标$\mathbf{C}_{\text{bone}}$。交叉注意力生成骨骼—顶点相关特征：

$$\mathbf{A}_{\text{rig}} = \text{CrossAttn}(\mathbf{B}^{\text{emb}}, \mathbf{V}_0^{\text{emb}}, \mathbf{V}_0^{\text{emb}}), \tag{2}$$

其中，每根骨骼都会预测高斯参数 $\mathbf{G}_k = [\Delta\mathbf{c}_k, \mathbf{s}_k, \mathbf{q}_k]$。

运动分支。运动分支对时间变形模式进行编码，并预测控制局部和全局运动的变分潜在变量。我们首先计算逐帧顶点位移：

$$\mathbf{V}_\Delta = \mathbf{V}[:, 1:, :, :] - \mathbf{V}[:, :-1, :, :] \in \mathbb{R}^{B\times(T-1)\times N\times 3}. \tag{3}$$

这些位移经过时空注意力层处理，得到 $\mathbf{V}_\Delta^{\text{emb}}$。利用共享的骨骼词元坐标 $\mathbf{C}_{\text{bone}}$，我们提取骨骼运动交互特征：

$$\mathbf{A}_{\text{motion}} = \text{CrossAttn}(\mathbf{B}^{\text{emb}}, \mathbf{V}_\Delta^{\text{emb}}, \mathbf{V}_\Delta^{\text{emb}}) \in \mathbb{R}^{B\times(T-1)\times K\times d_b}. \tag{4}$$

这些特征被输入局部运动后验估计器，为每根骨骼 k 和每一帧 t预测骨骼级后验参数：

$$[\boldsymbol{\mu}_{\text{local}}, \log \boldsymbol{\sigma}_{\text{local}}] = \text{MLP}(\mathbf{A}_{\text{motion}}). \tag{5}$$

随后通过以下方式采样潜在变量：

$$\mathbf{z}_{\text{local}} = \boldsymbol{\mu}_{\text{local}} + \boldsymbol{\sigma}_{\text{local}} \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}). \tag{6}$$

对于全局变换，全局运动后验估计器在时间维度上聚合运动特征（平均池化），生成根部级后验参数：

$$[\boldsymbol{\mu}_{\text{root}}, \log \boldsymbol{\sigma}_{\text{root}}] = \text{MLP}(\text{Agg}(\mathbf{A}_{\text{motion}})). \tag{7}$$

根部级潜编码采样如下：

$$\mathbf{z}_{\text{root}} = \boldsymbol{\mu}_{\text{root}} + \boldsymbol{\sigma}_{\text{root}} \odot \boldsymbol{\epsilon}. \tag{8}$$

3.1.2. 绑定运动解码器

解码器将潜编码映射为具有物理意义的变换，包括一个静态高斯骨骼解码器，以及用于局部和根部 SE(3) 变换的运动解码器。绑定分支生成的潜在特征作为输入，由一个轻量级的基于多层感知机的模块预测每根骨骼对应的软高斯骨骼区域 $\mathbf{G} = [\Delta\mathbf{c},\ \mathbf{s},\ \mathbf{q}]$；而局部运动和全局运动则根据运动编码器预测的潜变量进行解码。具体而言，局部运动解码器将 $\mathbf{z}_{\text{local}}$ 映射为每根骨骼的变换：

$$\{\mathbf{q}_{\text{local}}, \mathbf{t}_{\text{local}}\} = \text{Dec}_{\text{local}}(\mathbf{z}_{\text{local}}). \tag{9}$$

另一个根运动解码器将 $\mathbf{z}_{\text{root}}$ 映射为全局变换：

$$\{\mathbf{q}_{\text{root}}, \mathbf{t}_{\text{root}}\} = \text{Dec}_{\text{root}}(\mathbf{z}_{\text{root}}). \tag{10}$$

3.1.3. 高斯蒙皮线性混合蒙皮模块

高斯骨骼定义。每根高斯骨骼 k 均由其几何参数 $\mathbf{G}_k = [\mathbf{c}_k, \mathbf{s}_k, \mathbf{q}_k]$定义，其中$\mathbf{c}_k \in \mathbb{R}^3$ 表示骨骼中心，$\mathbf{s}_k \in \mathbb{R}^3$ 表示各向异性缩放因子，而$\mathbf{q}_k \in \mathbb{R}^4$ 是方向四元数。这些参数共同定义了一个三维高斯椭球，它充当软骨骼，具有随空间变化的影响范围。

蒙皮权重计算。对于每个顶点$\mathbf{v}_i$，使用高斯骨骼坐标系中的马氏距离计算其相对于骨骼 k 的蒙皮权重：

$$w_{ik}^{\text{raw}} = \frac{\exp\left(-\frac{1}{2}\|\mathbf{R}_k^\top(\mathbf{v}_i - \mathbf{c}_k) \oslash \mathbf{s}_k\|^2\right)}{\sum_{j=1}^{K}\exp\left(-\frac{1}{2}\|\mathbf{R}_j^\top(\mathbf{v}_i - \mathbf{c}_j) \oslash \mathbf{s}_j\|^2\right)}, \tag{11}$$

其中，$\mathbf{R}_k$ 是由四元数 $\mathbf{q}_k$导出的旋转矩阵，$\oslash$ 表示逐元素除法。

线性混合蒙皮。给定蒙皮权重和骨骼变换后，通过线性混合蒙皮对每个顶点进行变形：

$$\hat{\mathbf{v}}_i = \sum_{k=1}^{K} w_{ik} \cdot \mathbf{T}_k \cdot \tilde{\mathbf{v}}_i, \tag{12}$$

其中，$\mathbf{T}_k = \mathbf{T}_{\text{root}} \cdot \mathbf{T}_{k,\ \text{local}}$ 表示结合根运动与局部骨骼运动的层级变换，$\tilde{}\mathbf{v}_i$ 是齐次坐标中的顶点。

通过测地线感知权重细化实现拓扑一致性。虽然高斯加权能够捕捉局部邻近关系，但可能会错误地关联空间上接近、拓扑上相距较远的区域（例如手臂与躯干接触）。为减少此类伪影，我们引入一种测地线感知权重细化策略，以确保网格表面上的拓扑一致性。

设 $V = \{v_1, \ldots, v_N\}$ 表示顶点，$A = \{a_1, \ldots, a_K\}$ 表示具有原始权重 $W^{\text{raw}} \in \mathbb{R}^{N\times K}$ 的骨骼锚点。顶点 v_i 与锚点 a_k 之间的表面测地距离定义为：

$$d_g(v_i, a_k) = \min_{\pi\in\Pi(a_k, v_i)} \sum_{(v_p, v_q)\in\pi} \|v_p - v_q\|_2, \tag{13}$$

其中，$\Pi(a_k,\ v_i)$ 枚举边连通顶点路径。二值一致性掩码定义为

$$M_{ik} = \begin{cases} 1, & d_g(v_i, a_k) < \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

细化后的蒙皮权重为：

$$\tilde{W}_{ik} = W_{ik}^{\text{raw}} M_{ik}, \quad w_{ik} = \frac{\tilde{W}_{ik}}{\sum_{j=1}^{K}\tilde{W}_{ij} + \varepsilon}, \tag{15}$$

以确保数值稳定性。无法到达任何骨骼的顶点会被赋予一个指向最近骨骼的独热权重。这种优化有效抑制了跨部位影响，并产生更干净、更连贯的蒙皮，通常每个顶点仅受 2–3 根骨骼的影响。

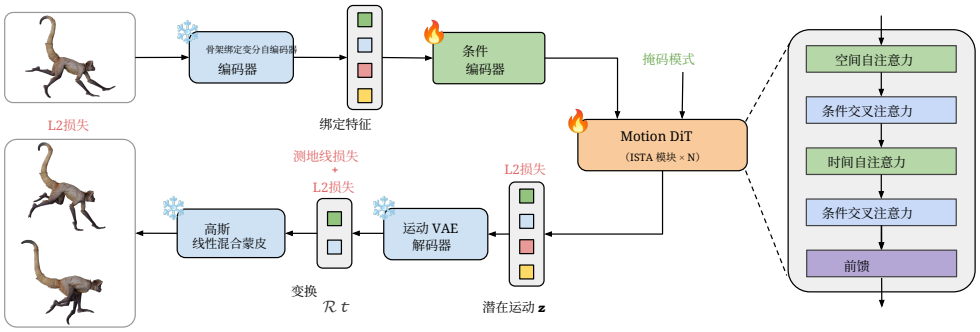


图3. **Motion DiT** 概览。给定静态骨架绑定特征，条件编码器会生成锚点令牌和全局令牌，用于引导扩散变换器在 RigMo 的运动潜在空间中运行。模型通过空间、时间和帧条件交叉注意力预测去噪后的运动潜变量，再通过高斯蒙皮将其解码为骨骼变换和顶点序列。

自监督训练目标。 RigMo 的一项关键优势在于，训练过程中无需骨架绑定标注或监督。相反，模型仅凭运动数据，以纯粹的自监督方式直接学习底层骨架结构。这省去了专业艺术家进行昂贵手工骨架绑定标注的需求，使该方法能够轻松扩展至大规模运动数据集。RigMo-VAE 仅通过两个自监督目标进行端到端训练：顶点级重建损失和潜变量正则化项：

$$\mathcal{L}_{\text{total}} = \lambda_{\text{recon}}\mathcal{L}_{\text{recon}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}}. \tag{16}$$

重建项对逐顶点运动保真度进行监督，而无需任何骨架绑定监督：

$$\mathcal{L}_{\text{recon}} = \frac{1}{BTN}\sum_{b,t,i}\|\hat{\mathbf{v}}_{b,t,i} - \mathbf{v}_{b,t,i}\|^2, \tag{17}$$

而 KL 散度则将潜在分布正则化至单位高斯先验：

$$\mathcal{L}_{\text{KL}} = \frac{1}{2}\sum_{i}(\mu_i^2 + \sigma_i^2 - \log\sigma_i^2 - 1). \tag{18}$$

这种纯粹由运动驱动的表述，使模型能够仅通过观察顶点轨迹，发现语义上有意义的骨骼结构及其关系，从而绕过传统手动绑定标注这一瓶颈，并支持从多样化的运动集合中进行可扩展学习。

3.2. Motion DiT

给定作为条件的静态绑定特征，Motion DiT 会在 RigMo-VAE 空间中生成运动潜变量。条件编码器将静态/绑定线索（例如锚点/高斯/蒙皮特征）聚合为锚点令牌 $\mathbf{A} \in \mathbb{R}^{B\times K\times H}$ 和全局令牌 $\mathbf{g} \in \mathbb{R}^{B\times H}$，这些令牌在生成过程中保持固定，并引导去噪过程。我们首先将变分自编码器的动态令牌 $\mathbf{Z}_{\text{dyn}} \in \mathbb{R}^{B\times K\times(T-1)\times D_{\text{dyn}}}$ 和根令牌 $\mathbf{Z}_{\text{root}} \in \mathbb{R}^{B\times(T-1)\times D_{\text{root}}}$ 投影到统一宽度 H，并通过拼接 K 骨骼流与根流，形成统一的运动潜变量张量 $\mathbf{Z}_{\text{mot}} \in \mathbb{R}^{B\times(K+1)\times(T-1)\times H}$。扩散变换器根据可配置的帧掩码调度（指定观测帧和生成帧）进行条件化，预测生成帧的速度场 $\hat{\mathbf{v}}_{\text{mot}}$，同时将观测帧固定为条件信号。随后通过 v 预测恢复 $\hat{\mathbf{x}}_0$，再由 RigMo-VAE 解码器解码为动画。这种以静态信息为条件、在运动潜变量中进行生成的设计，使合成动态与学习到的绑定结构紧密契合。

骨干网络由 $L{=}12$ 个交错的时空注意力（ISTA）块组成，隐藏维度为 $H{=}512$。每个块交替执行空间注意力（帧内、跨骨骼）和时间注意力（单个骨骼内、跨帧），并辅以两条条件通路：（1）通过交叉注意力注入静态和全局先验（$\mathbf{A}$，$\mathbf{g}$）；（2）帧级交叉注意力，根据掩码将生成帧与观测帧对齐。网络输出速度参数化 $\hat{\mathbf{v}}$，并据此重建去噪后的潜在状态 $\hat{\mathbf{x}}_0 = \sqrt{\alpha_t}\mathbf{x}_t - \sqrt{1-\alpha_t}\hat{\mathbf{v}}$。损失仅在生成帧上计算：

$$\mathcal{L} = \lambda_{\text{lat}}\mathcal{L}_{\text{lat}} + \lambda_{\text{rot}}\mathcal{L}_{\text{rot}} + \lambda_{\text{trans}}\mathcal{L}_{\text{trans}} + \lambda_{\text{vert}}\mathcal{L}_{\text{vert}},$$

其中，$\mathcal{L}_{\text{lat}}$ 表示潜在空间 L2 损失，$\mathcal{L}_{\text{rot}}$ 表示 SO(3) 测地旋转损失，$\mathcal{L}_{\text{trans}}$ 表示平移 L2 损失，$\mathcal{L}_{\text{vert}}$ 表示顶点空间 L2 损失。权重设为 $(\lambda_{\text{lat}},\ \lambda_{\text{rot}},\ \lambda_{\text{trans}},\ \lambda_{\text{vert}}) = (0.5,\ 1.0,\ 0.2,\ 0.1)$。这种设计使 Motion DiT 能够在 RigMo 的潜在空间中生成或插值复杂的长程运动，同时保持空间连贯性和时间真实感。

4. 实验

本文的主要重点是我们提出的 RigMo-VAE，它构成了绑定–运动表示的核心。我们还引入了 Motion DiT，以展示 RigMo-VAE 作为潜在

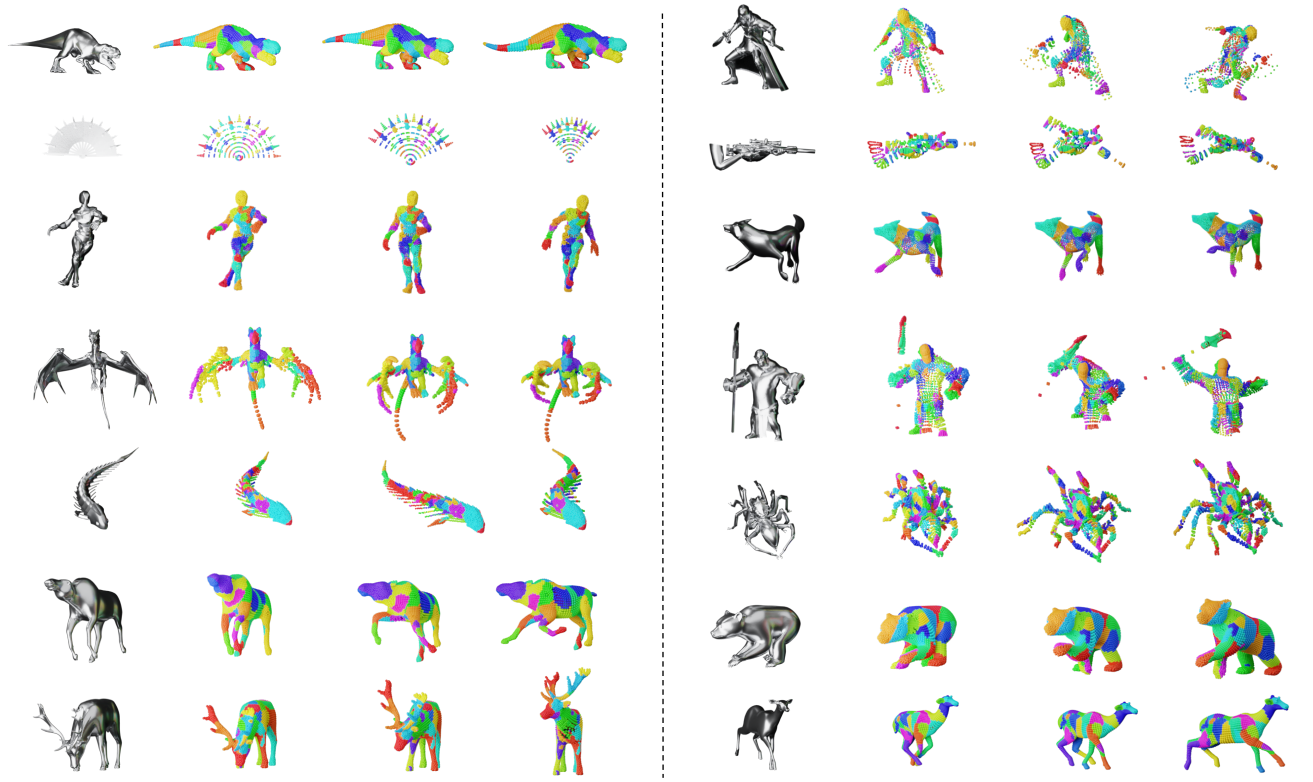


图4。完整 RigMo 生成的结果。给定一个稀疏输入序列，其中根据帧掩码观察部分帧，RigMo 通过联合预测绑定结构（高斯骨骼和蒙皮权重），并在 RigMo潜在空间中通过扩散合成缺失的运动帧，重建出完整的可动画模型。生成的绑定模型能够在人体、动物和各种非人类形状上产生连贯的关节运动，表明稀疏观测足以在不依赖类别特定先验的情况下恢复完整动画。

空间在下游运动生成中的有效性，但本工作的主要贡献在于变分自编码器本身的绑定与运动分解能力。受篇幅限制，Motion DiT 的更多结果请见补充材料。

4.1. 数据集

我们整理了一个大规模语料库，包含 ∼20,000 个可变形网格序列，涵盖三个互补领域，以确保运动覆盖的多样性和强泛化能力。**DeformingThings4D** [18]提供了 1,972 个真实世界序列，记录有机非刚性变形；为确保时间稳定性，我们排除了短于 10 帧的序列。 **TrueBones** [14]贡献了 1,287 个高保真关节动画，展现了逼真的骨骼运动。 **Objaverse-XL** [7]提供了 17,024 个合成序列（每个序列 24 帧），并通过运动质量分类器筛选，以保留不同类别中多样的拓扑结构和运动风格。对于每个数据集，我们采用 5:1 的比例划分训练集和测试集，以便在不同领域进行一致的评估。

为处理不同网格分辨率并保-持几何一致性，我们采用拓扑保持顶点归一化方案。对于顶点数量超过 20K 的网格，使用带有测地邻域保持的最远点采样（FPS）将其下采样至 5K；对于分辨率较低的网格，则迭代细分，直至顶点数量超过目标值，然后以相同方式下采样。该流程生成具有一致局部连通性的统一顶点嵌入，有助于实现稳健且具备拓扑感知能力的骨骼绑定发现。

4.2. 实现细节

RigMo-VAE 使用 PyTorch 实现，并在 24×A100（80GB）GPU 上以混合精度训练。每个批次包含 144 个网格序列（T=20, N=5K）。我们使用 AdamW（β_1=0.9, β_2=0.999, 权重衰减 10^{-4}），学习率为 3×10^{-4}，并采用 2K 步预热、余弦衰减、梯度裁剪（1.0）和指数移动平均（0.999）。编码器包含六个拓扑感知注意力层（隐藏维度 256，八个注意力头，邻域大小为 k=5）。锚点采样选取 K=48/128个高斯骨骼，并将蒙皮权重稀疏化，仅保留排名前 K_s=4 的权重。训练采用重建损失和 KL 损失（λ_{recon}=1.0, λ_{KL}=10^{-6}），并在训练前 30% 的过程中逐步增加 KL 损失的权重。在完整语料上，RigMo

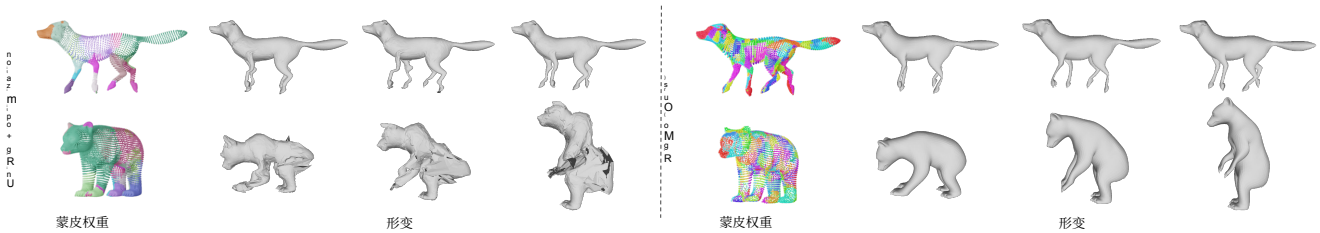


图5. UniRig+优化与我们的 RigMo绑定模块的比较。尽管 UniRig 在某些情况下（例如狐狸）可能会生成视觉上合理的蒙皮权重，但其绑定无法泛化，在实际动画中会崩溃，导致严重的变形伪影。相比之下，RigMo直接从运动中学习稳健且可迁移的骨架结构，无需任何真实绑定监督，并能在多种姿势和动物物种中实现稳定、高保真的形变。

方法	训练重建		跨动作迁移		均值 ↓	
	倒角距离-L1 ↓	倒角距离-L2 ↓	倒角距离-L1 ↓	倒角距离-L2 ↓	倒角距离-L1 ↓	倒角距离-L2 ↓
逐案例优化	12.3 ± 0.2	8.21 ± 0.3	68.8 ± 6.7	43.5 ± 5.4	40.55	25.86
UniRig [52] +优化	37.3 ± 2.3	28.4 ± 2.1	48.6 ± 4.9	31.2 ± 4.6	42.95	29.80
MagicArticulate [31] +选项	43.1 ± 3.7	23.9 ± 3.2	53.4 ± 5.2	28.7 ± 3.9	48.25	26.30
RigMo（我们的方法）	**11.1 ± 0.6**	**7.64 ± 0.3**	**13.82 ± 0.49**	**11.83 ± 0.34**	**12.46**	**9.74**

表 1. 骨骼绑定发现与跨动作泛化（$\times 10^{-3}$）。数值越低越好。实验在 DT4D 测试数据集上进行。

经过 ∼50K 步（∼10 天）后收敛。仅使用 DeformingThings4D 训练时，在 ∼7.5K 个 epoch（∼1.5 天）内即可达到相近的重建效果。推理完全采用前馈方式，在 A100 GPU 上重建一个 20 帧、含 5K 个顶点的序列，每帧耗时∼40毫秒。

Motion DiT使用 PyTorch 实现，采用正弦时间步嵌入，并将 12 层 ISTA 主干网络应用于 512 维运动令牌。我们采用包含 1000 个步长的 DDPM 调度器和 scaled-linear β，使用v-prediction 参数化。训练采用 AdamW（β_1=0.9、 β_2=0.999，权重衰减为 0.01），初始学习率为 1×10^{-4}，并使用余弦退火，在 4 张 H200 GPU 上进行（批量大小为 64）。端到端训练耗时 ∼40 小时，并在 ∼5000 个周期内收敛。得益于其时空注意力设计以及 RigMo 运动潜在空间中的稀疏条件化，该模型无需辅助损失也能保持稳定。

4.3. 绑定评估

评估协议。 我们在 DeformingThings4D 数据集划分上开展全面的绑定评估，利用其无标注特性评估真正的无监督绑定发现能力。实验协议随机从同一对象的训练数据和测试数据中抽取运动序列对，形成用于评估的训练集和测试集；训练序列用于基线优化，测试序列用于评估跨运动绑定迁移能力。我们通过 100 次独立随机划分确保统计显著性，并报告平均指标和置信区间。

竞争基线。 如表1所示，我们针对代表当前实践的两种典型方法进行了严格比较：**逐案例优化**。该基线通过对重建损失 $\mathcal{L}_{\text{recon}}$进行梯度下降，联合优化绑定参数{$\mathbf{G}$，$\mathbf{W}$} 和运动参数 {$\mathbf{T}_t$}。测试期间，绑定参数保持固定，仅优化变换，从而直接衡量泛化能力。然而，由于预测的绑定完全由训练序列本身决定，该方法只能在其针对特定案例进行优化时表现良好。应用于未见过的运动时，优化后的绑定往往会崩溃或变得无效，暴露出其泛化能力不足。 **自动绑定流程。**我们采用最先进的自动绑定方法（UniRig [52], MagicArticulate [31]），从规范姿态生成初始骨骼结构，然后针对每个序列优化变换。该流程体现了当前自动化角色绑定的行业标准。

如图5所示，自动绑定流程的泛化能力也受到训练数据多样性和结构的限制。在许多情况下，这些方法生成的骨骼布局在视觉上看似合理，但在动画过程中甚至无法复现常见的运动。重要的是，即使是经验丰富的艺术家，也无法仅凭视觉先验来分配正确的蒙皮权重；为了生成能够准确复现目标形变的绑定，始终需要进行迭代调整。这一观察与 RigMo 背后的核心动机一致——无法仅从静态几何中推断出稳健的绑定，而必须直接从运动中学习。

4.4. RigMo-VAE 重建评估

我们在精心筛选的子测试集上，与领先的网格生成架构进行了全面的重建基准测试。该测试集包含 500 个序列，涵盖多种形变

方法	几何精度 倒角距离-L1 ↓	倒角距离-L2 ↓	时间（20帧） ↓
AnimateAnyMesh [38]	1.81 ± 0.13	1.32 ± 0.01	2.8s
Step1X3D [17]	3.63 ± 0.21	2.96 ± 0.18	22.6秒
混元3D [54]	3.21 ± 0.19	2.67 ± 0.14	17.4秒
RigMo（我们的方法）	**1.73 ± 0.11**	**1.26 ± 0.08**	**0.74秒**

表 2. 重建保真度与推理效率（$\times 10^{-2}$ 倒角距离）。RigMo 在所有基线方法中实现了最高的几何精度和最快的推理速度。

运动模式和物体类别。

我们将其与兼顾序列感知和逐帧独立生成模型的代表性范式进行比较（见表2）：**AnimateAnyMesh** [38]采用序列感知的变分自编码器架构来建模时序网格。然而，其设计需要 512 个令牌的高维表示来编码物体运动，而 RigMo-VAE 仅用 48/128 个令牌便实现了更优的重建质量，从而显著加快推理速度并减少内存占用。**Step1X3D** [17] 和**Hunyuan 3D 2。1** [54] 代表了当前最先进的逐帧独立三维变分自编码器，它们逐帧生成，因此在序列重建时会带来巨大的计算开销。这些方法本质上会破坏不同帧之间的顶点对应关系，并且由于采用逐帧解码，往往会丢失精细的表面细节。由于它们的潜在空间基于不一致的规范化，我们的评估采用了姿态不变校正：记录用于输入归一化的规范变换 $\{\mathbf{R},\ \mathbf{t}\ s\}$ 并在重建后应用其逆变换，以便进行公平且聚焦于形状的几何评估。

4.5. 定性分析

如图所示。4，我们从三个关键方面展示了全面的可视化结果，体现了 RigMo 的语义理解能力和生成能力：**学习得到的绑定结构**。可视化结果揭示了无需监督即可形成的、具有语义意义的骨骼-顶点对应关系；在不同物体类别中，不同的高斯骨骼能够捕捉直观的解剖区域（四肢、躯干和末端部位）。**可控运动合成**。Motion DiT 的演示展现了精确的条件生成能力，能够生成合理的动画，在遵循学习得到的绑定约束的同时，支持对新颖运动模式进行创造性探索。

4.6. 诊断分析

分辨率无关架构。无论原始分辨率如何，所有网格都会先重采样为5000顶点表示，再由编码器处理。由于高斯骨骼和运动变换是在三维空间中连续定义的，而非绑定到特定的顶点索引，

配置	倒角距离-L1 ↓	倒角距离-L2 ↓
不含测地线细化	2.37 ± 0.15	2.07 ± 0.18
48 个骨骼令牌	1.91 ± 0.13	1.48 ± 0.12
128 个骨骼令牌	**1.73** ± **0.11**	**1.26** ± **0.08**

表3. 在 DeformingThings4D 验证集上使用 CD-L1/L2 指标进行的消融研究（$\times 10^{-2}$）。

预测得到的绑定和运动参数可以直接应用回原始分辨率的网格。即使网格剖分发生变化，这一特性也能保持变形质量一致，因为学习到的骨骼捕捉的是几何区域，而非固定的顶点子集。这与依赖网格的方法形成对比： [38], 在这类方法中，顶点数量的变化通常会改变学习到的蒙皮结构，并降低动画保真度。

拓扑细化的影响。 接下来，我们评估测地线感知权重细化模块的有效性。如表3所示，移除这一组件会导致重建质量大幅下降。这些结果表明，对于关节式物体而言，仅依靠空间邻近性并不足够，因为顶点在欧几里得空间中可能彼此接近，但在底层运动学拓扑中却相距甚远。

骨骼令牌数量权衡。 最后，我们分析骨骼令牌数量的影响。将令牌数从 48 增加到 128，CD-L1 提升了 0.018%；但随着令牌变得过于细粒度，收益会逐渐减小。尽管 128 个令牌取得了最佳的定量重建效果，48 个令牌在效率、可解释性和稳定性之间实现了更理想的平衡。过多的令牌往往会割裂连贯的解剖区域，却无法带来相称的性能提升。

5. 结论

本文提出 RigMo，这是一种统一的生成式框架，可在无需任何骨架标注的情况下，直接从原始网格序列中联合学习骨架结构与运动动力学。RigMo 将逐顶点变形分解到两个紧凑的潜在空间中，并将其解码为显式的高斯骨骼、蒙皮权重和骨骼变换，从而在构造层面生成完全可动画化的表示。结合在这些结构感知潜变量上运行的 Motion-DiT 模型，RigMo 实现了可控的运动生成及无缝的下游应用。在多个数据集上的实验表明，RigMo 能够发现平滑、可解释且符合物理规律的绑定，同时在准确性和泛化能力方面优于现有的自动绑定与形变方法。

参考文献

[1] Kfir Aberman、Peizhuo Li、Dani Lischinski、Olga Sorkine-Hornung、Daniel Cohen-Or 和 Baoquan Chen。用于深度动作重定向的骨架感知网络。ACM Transactions on Graphics (TOG)，39(4):62–1，2020。2[2] Sherwin Bahmani、Xian Liu、Wang Yifan、Ivan Skorokhodov、Victor Rong、Ziwei Liu、Xihui Liu、Jeong Joon Park、Sergey Tulyakov、Gordon Wetzstein 等。TC4D：轨迹条件文本到4D生成。载于 ECCV，2024。2[3] Sherwin Bahmani、Ivan Skorokhodov、Victor Rong、Gordon Wetzstein、Leonidas Guibas、Peter Wonka、Sergey Tulyakov、Jeong Joon Park、Andrea Tagliasacchi 和 David B Lindell。4D-FY：使用混合分数蒸馏采样的文本到4D生成。载于 CVPR，2024。2[4] Ilya Baran 和 Jovan Popovi´c。3D角色的自动绑定与动画。载于 ACM TOG，2007。2[5] Wei Cao、Chang Luo、Biao Zhang、Matthias Nießner 和 Jiapeng Tang。Motion2vecsets：用于非刚性形状重建与跟踪的4D潜向量集扩散。载于 IEEE/CVF计算机视觉与模式识别会议，页码 20496–20506，2024。2[6] Jianqi Chen、Biao Zhang、Xiangjun Tang 和 Peter Wonka。V2M4：从单目视频进行4D网格动画重建。arXiv 预印本，arXiv:2503.09631，2025。2[7] Matt Deitke、Ruoshi Liu、Matthew Wallingford、Huong Ngo、Oscar Michel、Aditya Kusupati、Alan Fan、Christian Laforte、Vikram Voleti、Samir Yitzhak Gadre 等。Objaverse-XL：一个包含1000万+ 3d对象的宇宙。NeurIPS，2023。2，6[8] Zhoujie Fu、Jiacheng Wei、Wenhao Shen、Chaoyue Song、Xiaofeng Yang、Fayao Liu、Xulei Yang 和 Guosheng Lin。Sync4d：用于基于物理的4D生成的视频引导可控动力学。arXiv 预印本，arXiv:2405.16849，2024。2[9] Inbar Gat、Sigal Raab、Guy Tevet、Yuval Reshef、Amit Haim Bermano 和 Daniel Cohen-Or。AnyTop：具有任意拓扑的角色动画扩散。载于图形学与交互技术特别兴趣小组会议论文集，页码 1–10，2025。2。[10] Jonathan Ho、Ajay Jain 和 Pieter Abbeel。去噪扩散概率模型。神经信息处理系统进展，33:6840–6851，2020。2[11] Seokhyeon Hong、Chaelin Kim、Serin Yoon、Junghyun Nam、Sihun Cha 和 J unyong Noh。SALAD：用于文本驱动动作生成与编辑的骨架感知潜扩散。载于计算机视觉与模式识别会议论文集，页码 7158–7168，2025。2[12] Yanqin Jiang、Li Zhang、Jin Gao、Weimin Hu 和 Yao Yao。Consistent4d：一致的 360 {\deg} 动态对象生成。载于单目视频。arXiv 预印本，arXiv:2311.02848，2023。2[13] Yanqin Jiang、Chaohui Yu、Chenjie Cao、Fan Wang、Weiming Hu 和 Jin Gao。Animate3d：利用多视角视频扩散为任意3D模型制作动画。NeurIPS，2024。2

[14] Wonkwang Lee、Jongwon Jeong、Taehong Moon、Hyeon- Jong Kim、Jaehyeon Kim、Gunhee Kim 和 Byeong-Uk Lee。《如何让你的龙动起来：面向大词汇量物体的文本到动作合成》。arXiv 预印本 arXiv:2503 04257，2025 6。.[15] Peizhuo Li、Kfir Aberman、Rana Hanocka、Libin Liu、Olga Sorkine-Hornung 和 Baoquan Chen。《利用神经混合形状学习骨骼关节运动》。ACM TOG，2021。2[16] Weiyu Li、Xuelin Chen、Peizhuo Li、Olga Sorkine-Hornung 和 Baoquan Chen。《通过生成式动作匹配进行基于示例的动作合成》。ACM图形学汇刊（TOG），42(4):1–12，2023。2[17] Weiyu Li、Xuanyang Zhang、Zheng Sun、Di Qi、Hao Li、Wei Cheng、Weiwei Cai、Shihao Wu、Jiarui Liu、Zihao Wang 等。《Step1x-3d：迈向高保真、可控的纹理化三维资产生成》。arXiv 预印本 arXiv:2505.07747，2025。8[18] Yang Li、Hikari Takehara、Takafumi Taketomi、Bo Zheng 和 Matthias Nießner。《4DComplete：超越可观测表面的非刚性运动估计》。发表于 ICCV，2021。2, 6[19] Zhiqi Li、Yiming Chen 和 Peidong Liu。《Dreammesh4d：采用稀疏控制高斯-网格混合表示的视频到4D生成》。神经信息处理系统进展，37:21377–21400，2024。2[20] Zhouyingcheng Liao、Jimei Yang、Jun Saito、Gerard Pons-Moll 和 Yang Zhou。《风格化3D角色的无骨架姿态迁移》。发表于 ECCV，第640–656页，2022。1[21] Isabella Liu、Zhan Xu、Wang Yifan、Hao Tan、Zexiang Xu、Xiaolong Wang、Hao Su 和 Zifan Shi。《RigAnything：面向多样化3D资产的无模板自回归绑定》。ACM TOG，2025。2[22] Matthew Loper、Naureen Mahmood、Javier Romero、Gerard Pons-Moll 和 Michael J Black。《SMPL：一种蒙皮多人物线性模型》。发表于《奠基性图形学论文：拓展边界》，第 *2*， 851 866 2023 2 页码 –。 . Ji[23] ng Ma 和 Dongliang Zhang。《TARIG：面向类人角色的自适应模板感知神经网络骨骼绑定》。计算机与图形学，2023。2[24] Ekkasit Pinyoanuntapong、Muhammad Saleem、Korrawe Karunratanakul、Pu Wang、Hongfei Xue、Chen Chen、Chuan Guo、Junli Cao、Jian Ren 和 Sergey Tulyakov。《Maskcontrol：掩码动作合成的时空控制》。发表于 IEEE/CVF国际计算机视觉会议论文集，第9955–9965页，2025。2[25] Sigal Raab、Inbal Leibovitch、Peizhuo Li、Kfir Aberman、Olga Sorkine-Hornung 和 Daniel Cohen-Or。《MODI：从多样化数据中进行无条件动作合成》。发表于 IEEE/CVF计算机视觉与模式识别会议论文集，第13873–13883页，2023。2[26] Sigal Raab、Inbal Leibovitch、Guy Tevet、Moab Arar、Amit H Ber mano 和 Daniel Cohen-Or。《单动作扩散》。arXiv 预印本 arXiv:2302.05905，2023。2[27] Jiawei Ren、Cheng Xie、Ashkan Mirzaei、Karsten Kreis、Zi- wei Liu、Antonio Torralba、Sanja Fidler、Seung Wook Kim、Huan Ling 等。《L4GM：大型4D高斯重建模型》。神经信息处理系统进展，37:56828–56858，2024。2

[28] Yahao Shi、Yang Liu、Yanmin Wu、Xing Liu、Chen Zhao、Jie Luo 和 Bin Zhou。《驱动任意网格：基于视频的网格变形4D潜在扩散》。arXiv 预印本 arXiv:2506.07489，2025。2[29] Chaoyue Song、Jiacheng Wei、Ruibo Li、Fayao Liu 和 Gu- osheng Lin。《通过对应关系学习与网格细化进行3D姿态迁移》。发表于 NeurIPS，第3108 3120页，2021 1 –。[30] Chaoyue Song、Xiu Li、Fan Yang、Zhongcong Xu、Jiacheng Wei、Fayao Liu、Jiashi Feng、Guosheng Lin 和 Jianfeng Zhang。《Puppeteer：为你的3D模型绑定并制作动画》。NeurIPS，2025。2[31] Chaoyue Song、Jianfeng Zhang、Xiu Li、Fan Yang、Yiwen Chen、Zhongcong Xu、Jun Hao Liew、Xiaoyang Guo、Fayao Liu、Jiashi Feng 等。《Magicarticulate：让你的3D模型可直接进行关节运动》。发表于 CVPR，2025。2, 7[32] Keqiang Sun、Dor Litvak、Yunzhi Zhang、Hongsheng Li、Ji- aj un Wu 和 Shangzhe Wu。《Ponymation：从未标注网络视频中学习3D动物的关节运动》。发表于欧洲计算机视觉会议，第100–119页。施普林格，2024。2[33] Guy Tevet、Sigal Raab、Brian Gordon、Yonatan Shafir、Daniel Cohen-Or 和 Amit H Bermano。《人体运动扩散模型》。arXiv 预印本 arXiv:2209.14916，2022。2[34] Shashank Tripathi、O mid Taheri、Christoph Lassner、Michael Black、Daniel Holden 和 Carsten Stoll。《Humos：以身体形状为条件的人体运动模型》。发表于欧洲计算机视觉会议，第133–152页。施普林格，2024。2[35] Xinzhou Wang、Yikai Wang、Junliang Ye、Fuchun Sun、Zhengyi Wang、Ling Wang、Pengkun Liu、Kai Sun、Xin- tong Wang、Wende Xie 等。《Animatabledreamer：通过规范评分蒸馏进行文本引导的非刚性3D模型生成与重建》。发表于欧洲计算机视觉会议，第321–339页。施普林格，2024。2[36] Shangzhe Wu、Ruining Li、Tomas Jakab、Christian Rup- precht 和 Andrea Vedaldi。《Magicpony：在野外环境中学习关节运动的3D动物》，2023。2[37] Zijie Wu、Chaohui Yu、Yanqin Jiang、Chenjie Cao、Fan Wang 和 Xiang Bai。《Sc4d：稀疏控制的视频到4D生成与运动迁移》。发表于欧洲计算机视觉会议，第361–379页。施普林格，2024。2[38] Zijie Wu、Chaohui Yu、Fan Wang 和 Xiang Bai。《Animateanymesh：用于文本驱动通用网格动画的前馈式4D基础模型》。ICCV，2025。2, 8[39] Yiming Xie、Chun-Han Yao、Vikram Voleti、Huaizu Jiang 和 Varun Jampani。《Sv4d：兼具多帧与多视图一致性的动态3D内容生成》。arXiv 预印本 arXiv:2407.17470，2024。2[40] Zhan Xu、Yang Zhou、Evangelos Kalogerakis、Chris Landreth 和 Karan Singh。《RigNet：面向关节运动角色的神经网络骨骼绑定》。arXiv 预印本 arXiv:2005.00559，2020。2[41] Sijie Yan、Yuanjun Xiong 和 Dahua Lin。《用于基于骨架的动作识别的时空图卷积网络》。发表于 AAAI人工智能会议论文集，2018。2

[42] Gengshan Yang、Deqing Sun、Varun Jampani、Daniel Vlasic、Forrester Cole、Huiwen Chang、Deva Ramanan、William T. Freeman 和 Ce Liu。Lasr：从单目视频中学习关节形状重建，2021。2[43] Chun-Han Yao、Yiming Xie、Vikram Voleti、Huaizu Jiang 和 Varun Jampani。Sv4d 2.0：在多视角视频扩散中增强时空一致性，实现高质量4D生成。arXiv预印本 arXiv:2503.16396，2025。2[44] Ye Yuan、Jiaming Song、Umar Iqbal、Arash Vahdat 和 Jan Kautz。Physdiff：物理引导的人体运动扩散模型。载于IEEE/CVF国际计算机视觉会议论文集，第16010–16021页，2023。2[45] Yifei Zeng、Yanqin Jiang、Siyu Zhu、Yuanxun Lu、Youtian Lin、Hao Zhu、Weiming Hu、Xun Cao 和 Yao Yao。Stag4d：时空锚定生成式4D高斯。载于欧洲计算机视觉会议，第163–179页。施普林格，2024。2[46] Bowen Zhang、Sicheng Xu、Chuxin Wang、Jiaolong Yang、Feng Zhao、Dong Chen 和 Baining Guo。用于高保真视频到4D合成的高斯变分场扩散。载于ICCV，2025。2[47] Hao Zhang、Di Chang、Fang Li、Mohammad Soleymani 和 Narendra Ahuja。Magicpose4d：通过外观与运动控制构建关节模型。arXiv预印本 arXiv:2405.14017，2024。2[48] Haiyu Zhang、Xinyuan Chen、Yaohui Wang、Xihui Liu、Yunhong Wang 和 Yu Qiao。4diffusion：用于4D生成的多视角视频扩散模型。《神经信息处理系统进展》，37：15272–15295，2024。2[49] Hao Zhang、Fang Li、Samyak Rawlekar 和 Narendra Ahuja。S3o：从单目视频重建关节物体动态形状与骨架的双阶段方法。载于第41届国际机器学习会议论文集，第59191–59209页，2024。2[50] Hao Zhang、Fang Li、Samyak Rawlekar 和 Narendra Ahuja。学习隐式表示以重建关节物体。载于第十二届国际学习表示会议，2024。[51] Hao Zhang、Haolan Xu、Chun Feng、Varun Jampani 和 Narendra Ahuja。Physrig：用于真实铰接物体建模的可微物理蒙皮与绑定框架。ICCV，2025。2[52] Jia-Peng Zhang、Cheng-Feng Pu、Meng-Hao Guo、Yan-Pei Cao 和 Shi-Min Hu。一个模型绑定所有对象：使用UniRig进行多样化骨架绑定。arXiv预印本 arXiv:2504.12451，2025。2, 7[53] Mingyuan Zhang、Zhongang Cai、Liang Pan、Fangzhou Hong、Xinying Guo、Lei Yang 和 Ziwei Liu。Motiondif-fuse：使用扩散模型进行文本驱动的人体运动生成。载于IEEE TPAMI，2024。2[54]Zibo Zhao、Zeqiang Lai、Qingxiang Lin、Yunfei Zhao、Haolin Liu、Shuhui Yang、Yifei Feng、Mingxin Yang、Sheng Zhang、Xianghui Yang 等。混元3D 2.0：扩展扩散模型以生成高分辨率纹理3D资产。arXiv 预印本 arXiv:2501.12202，2025。8[55] Chongyang Zhong、Lei Hu、Zihao Zhang 和 Shihong Xia。Attt2m：使用

多视角注意力机制。载于*IEEE/CVF*国际计算机视觉会议论文集，第509–519页，2023。 2[56] Hanxin Zhu、Tianyu He、Xiqian Yu、Junliang Guo、Zhibo Chen 和 Jiang Bian。Ar4d：从单目视频进行自回归4D生成。*arXiv*预印本 *arXiv:2501 01722*，20252。。 [57] Qiran Zou、Shangyuan Yuan、Shian Du、Yu Wang、Chang Liu、Yi Xu、Jie Chen 和 Xiangyang Ji。Parco：部件协调的文本到动作合成。载于 欧洲计算机视觉会议，第126–143页。施普林格，2024。2[58] Silvia Zuffi、Angjoo Kanazawa、David W Jacobs 和 Michael J Black。3D动物园：动物的三维形状和姿态建模。载于 *IEEE*计算机视觉与模式识别会议论文集，第6365–6373页，20172–。

RigMo：统一绑定与运动学习的生成式动画

补充材料

补充材料概述

由于主论文的页数严格受限，若干重要细节、分析和可视化内容未能呈现。本文档对我们的方法和实验进行了全面扩展。具体而言，我们补充了以下内容：

- **与先前工作的比较**。
- **RigMo-VAE**。 我们并列展示了 48 个词元和 128 个词元配置，以突出它们在重建质量方面的差异。
- **MotionDiT**。我们详细介绍了实验设置、训练配置、数据集准备和完整的定量评估，并对观察到的趋势进行了深入分析。

•视频演示。 我们还提供了展示 RigMo 生成结果（1帧→ 9帧预测）的演示视频，以及与无法纳入主论文的当前最先进基线方法进行重建结果并排视觉比较的视频。

这些补充材料共同呈现了 RigMo 有效性更完整的图景，并支持主论文中的论断。

与先前工作的比较。

与 RigMo 相关的现有方法大致分为两类：动态三维生成方法，以及绑定—结构预测方法。尽管这些范式取得了令人瞩目的进展，但它们都未能以统一的自监督方式联合建模绑定与运动，这从根本上限制了其对一般可变形物体的适用性。

动态三维生成。 近期的四维VAE模型，如 AnimateAnyMesh 和 GVFDiffusion，会学习每个顶点或高斯基元的轨迹。尽管这些模型在短期序列重建方面表现有效，但无法恢复显式的绑定或运动学结构，而且通常需要数百个潜在标记（通常为 512+）来编码运动。相比之下，RigMo 仅使用 48 个潜在标记，却能生成紧凑、可解释的绑定，以及时间连贯的运动参数。另一类研究则将静态三维VAE（包括混元3D和Step1X3D）扩展到逐帧网格生成。这些方法针对彼此独立的帧进行处理，因此存在推理速度慢、缺乏时间平滑性、几何形状随时间变化不一致等问题，而最关键的是，生成的形状之间没有顶点对应关系。最后，运动生成

方法	ExplicitRig	时间一致性	骨架标注
动态三维生成			
AnimateAnyMesh / GVFDiff.	✗	✓	none
混元3D / Step1X3D	✗	✗	none
Human/AnyTop 运动生成	✓(GT)	✓	真实标注骨架
骨架预测			
UniRig / Magic Articul.	✓	✗	真实标注骨架
RigMo（我们的方法）			
RigMo	✓	✓	none

表4. RigMo 与代表性方法类别的比较。

模型专为人类或可关节运动的类别（例如 AnyTop，以及许多人类运动扩散/Transformer模型）设计，并假设可以获取真实标注骨架。对于不存在一致骨架标注的通用物体，这一假设便不再成立；此外，不同数据集和艺术家的绑定惯例差异很大，导致模型跨类别泛化能力较差。

骨架预测。 UniRig 和 Magic Articulate 等方法旨在根据静态几何预测骨架结构。这些方法严重依赖大规模人工标注的骨架数据集；然而，手动绑定存在不一致、标注噪声，以及不同艺术家和建模标准之间的显著差异。此外，现有数据集规模较小、针对特定类别且难以扩展，这限制了模型对多样物体和运动模式的泛化能力。相比之下，RigMo 无需人工标注，能够直接从观测到的运动中推断潜在的运动学结构。这种由运动驱动的建模方式能够生成具有物理意义、在不同实例间保持一致，并与物体实际形变行为自动对齐的绑定。

通过将绑定和运动统一到单个生成模型中，RigMo 克服了这两类方法的核心局限。它无需依赖人工标签，即可提供可解释的骨架、一致的网格形变、时间上稳定的轨迹以及强大的泛化能力，为通用物体动画提供了一种有原则且可扩展的解决方案。

MotionDiT 评估

为了更全面地评估 MotionDiT 的运动生成能力，我们设计了一组受控实验，以最大限度地减少采样随机性，并单独考察基于扩散的时序预测模块所发挥的作用。每项实验均以以下所述的两种互补信号为条件。

方法	1→1		1→9	
	训练 (L1/L2)	Val (L1/L2)	训练 (L1/L2)	Val (L1/L2)
无帧条件	2.13±0.46 / 1.52±0.33	2.48±0.48/ 1.74±0.35	2.63±0.94 / 1.98±0.72	3.10±0.91/ 2.22±0.69
潜在骨架条件	1.89±0.42 / 1.31±0.32	2.01±0.43/ 1.45±0.31	2.15±0.85 / 1.44±0.64	2.51±0.84/ 1.78±0.60
完整 MotionDiT	**1.43**±0.41 / **0.99**±0.30	**1.77**±0.41 / **1.23**±0.29	**1.61**±0.83 / **1.18**±0.61	**1.86**±0.82 / **1.37**±0.56

表5。在两种稀疏条件设置下对 MotionDiT ($\times 10^{-2}$) 进行的消融研究。我们报告训练集和验证集上的倒角距离（L1 / L2）。三个变体包括：(1) 不使用帧条件（无帧掩码引导），(2) 潜在骨架条件（仅使用骨架分支潜在特征），以及 (3) 完整 MotionDiT（潜在骨架特征 + 解码后的蒙皮权重 + 高斯骨骼中心 + 帧掩码引导）。实验在 DT4D 测试数据集上进行。

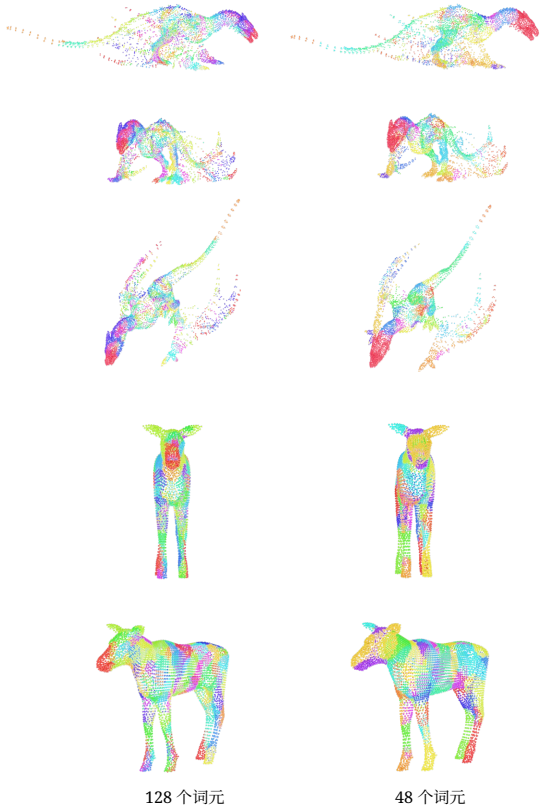


图6。48 个词元和128令牌配置的蒙皮权重并列对比。

• 骨架绑定条件：MotionDiT 接收从 RigMo-VAE 骨架分支中提取的、感知绑定的结构线索。这些线索包括潜在骨架特征，以及经过解码的、具有物理意义的绑定属性，例如蒙皮权重和高斯骨骼中心，并将其注入第一层交叉注意力层（图 3）。这种条件输入为模型提供了关节结构和预期的形变范围。

• 帧条件（掩码模式）：将一组稀疏的观测帧编码为帧掩码序列，并输入第二个交叉注意力层。此掩码模式控制模型可见的帧，并支持受控插值和长时域预测。

我们在两种稀疏条件设置下评估 MotionDiT：

1. 1 帧 $\rightarrow$ 1-帧预测：给定时间 t的单个输入帧，模型预测时间$t+1$的下一帧，重点关注局部时间平滑性和短期运动保真度。2. 1 帧 $\rightarrow$ 9-帧预测：给定一个观测帧，模型预测后续九帧。这种更具挑战性的设置用于评估模型生成时间连贯且符合物理规律的长时域运动的能力。

这些稀疏条件设置避免了自由形式运动采样，而是对 MotionDiT 的时间预测能力进行受控评估。它们还有助于清晰考察不同绑定条件策略的影响。

我们比较了三个变体：

• 无帧条件：完全移除帧掩码信号，迫使模型在没有任何观测帧引导的情况下推断时间结构。

• 潜在骨架条件：仅使用从 RigMo-VAE 绑定编码器提取的潜在骨架特征作为模型条件，不使用解码后的蒙皮权重或高斯骨骼中心。

•完整 MotionDiT：使用完整的骨架条件组合，包括潜在骨架特征、解码后的蒙皮权重和高斯骨骼中心，为运动预测提供明确且具有物理可解释性的先验。

在训练集和验证集划分上的定量结果，使用倒角距离（CL_1 和 CL_2）进行测量，汇总于表5中。在两个预测时域下，完整 MotionDiT 模型均展现出出色的运动保真度和时间连贯性。