arXiv:2601.04194v1 [cs.CV] 7 Jan 2026

编排动态物体的世界

YanzheLyu[1], *, † ChenGeng[1], * Karthik Dharmarajan[1]Yunzhi Zhang[1] Hadi Alzayer[1,3] Shangzhe Wu[2] Jiajun Wu[1]1斯坦福大学 2剑桥大学 3马里兰大学



图 1. **我们的方法生成的 4D场景运动**。 我们提出了 CHORD，这是一种通用生成管线，能够为包含**多个** 相互交互的物体的场景制作动画。项目页面：https://yanzhelyu.github.io/chord

摘要

我们物理*4D*（*3D* + 时间）世界中的动态物体不断演化、变形，并与其他物体发生交互，从而产生多样的*4D*场景动态。在本文中，我们提出了一种通用的生成管线，即CHORD，用于***CHOR***编排*Dynamic*物体和场景，并合成这类现象。用于创建这些动态的传统基于规则的图形管线依赖类别特定启发式方法，但劳动密集且难以扩展。近期的基于学习的方法通常需要大规模数据集，而这些数据集可能无法涵盖所有感兴趣的物体类别。相反，我们的方法通过提出一种基于蒸馏的管线，从二维视频的欧拉表示中提取其中隐藏的丰富拉格朗日运动信息。我们的方法具有通用性、灵活性，并且与类别无关。我们通过开展实验来生成多种多物体*4D*动态，展示其相较于现有方法的优势，并证明其在生成机器人操作策略方面的适用性，从而展示了该方法的有效性。项目页面：*https://yanzhelyu.github.io/chord*

*共同贡献。†此项工作是在 Y. Lyu 作为斯坦福大学访问学生期间完成的。Y. Lyu 目前就读于中国科学技术大学。

1. 引言

人类和其他具身智能体生活在一个4D（3D + 时间）世界中，这个世界由各种各样的动态物体组成，即能够演变、变形或与其他物体交互的物体。在构建3D世界时，为物体形变和交互创建4D运动至关重要

机器人学模型 [48, 74]和具身人工智能 [36]。

传统上，要为由动态物体静态快照组成的场景生成此类运动一直很有挑战性，因为这需要大量的手动建模和专家投入。近期方法 [73] 尝试以端到端的方式，纯粹从数据中学习此类4D生成器。然而，大多数现有数据集 [13]都侧重于单个物体的内部形变和演变，几乎没有涉及物体间的交互；同时，描述物体形变和物体交互的4D数据也极为稀少。场景级4D动态数据的匮乏，使得现有数据驱动方法只能生成单个物体的动态。

受近期通用视频生成模型取得成功的启发，我们采用不同的方法来解决这个问题：从视频生成模型中蒸馏场景运动。总体而言，我们迭代优化每个物体的低层拉格朗日形变。在每个优化步骤中，我们对3D场景进行形变，并从特定视点进行渲染，再让视频生成模型判断这种形变是否合理。通过这一过程，我们实际上将视频模型用作高层“编舞者”，规划各个物体的运动，并使它们彼此协调一致。

尽管这种基于蒸馏的范式前景可期，但要借此获得合理的结果仍然颇具挑战。现有方法 [4, 30, 64]主要在物体级进行操作，生成的运动往往会出现明显的伪影。有两个主要障碍妨碍这些方法在我们的场景中有效发挥作用：(1) 4D形变在空间上维度很高，在时间上又缺乏充分约束；(2) 现代视频生成模型采用的非常规架构设计与现有蒸馏算法 [53]不兼容。

我们通过分析 4D 形变的内在局部性来应对第一个挑战：时间形变场在空间和时间上都应保持局部平滑。为此，我们设计了一种由粗到细的 4D 运动表示，在空间域和时间域中均引入层次结构。在空间上，我们采用双层基于控制点的表示，将精细的运动细节与粗粒度变换解耦。在时间上，受理论算法设计中一种历史悠久的数据结构——芬威克树 [15, 34], ——启发，我们以累积的基于区间的结构存储变形，从而隐式地保证时间一致性，并提升长时域运动的可学习性。凭借这两项创新，我们提出的新颖 4D 表示具有鲁棒性和稳定性，并支持生成多种多样的运动。

第二个挑战源于现代视频生成模型基于基于流的模型 [44]。这些模型与传统的蒸馏算法不兼容，因此，我们提出了一种从现代基于整流流的视频生成模型中进行蒸馏的新策略。我们为基于流的视频扩散模型推导出一种新颖的分数蒸馏采样（SDS） [53] 目标，并分析其噪声模式，从而使视频模型能够有效地为我们的 4D 表示提供指导。

通过提出这两项创新以及用于编排物体运动的框架，我们为在场景中生成与 4D 一致的动态物体运动这一难题，提供了一种简单而优雅的解决方案。我们将这套流程命名为“CHORD”，意为**编排动态**物体与场景。CHORD 通用、灵活，可应用于广泛的动态现象。我们在多种动态物体上评估了该框架，并将其与先前方法进行比较，结果显示出明显优势。

除了视觉生成之外，我们的流程还能够通过生成现实世界物体符合物理规律的拉格朗日形变轨迹，实现物理世界中的机器人操作。我们利用生成的3D轨迹规划真实机器人的运动，并展示这些轨迹能够引导机器人对多种动态物体进行零样本操作。

总之，我们的贡献如下：
1. 一种4D运动表示，将受芬威克树启发的累积时间结构与分层的从低到高自由度参数化相结合，非常适合基于蒸馏的四维生成。 2. 一种针对现代基于流的视频生成模型的蒸馏策略，使分数蒸馏采样算法能够从2D视频生成模型生成4D运动。 3. 一个稳健的框架，用于为多种动态物体生成符合物理规律的4D运动，并将其应用于学习现实世界中的机器人操作策略。

2. 相关工作

物体级4D生成。 生成具有4D一致性的物体变形一直是该领域长期面临的挑战。传统方法首先确定特定类别的运动学模型（即，绑定表示） [6, 8, 22, 43, 45, 51, 71, 88], 然后基于这些模型生成运动 [20, 28, 42, 47, 52, 58, 60, 63, 77],，这本质上将这些方法限制在有限的类别中。一些方法 [73, 82, 86] 尝试从现有的4D物体数据集 [12–14],中端到端地学习4D生成器，但由于现有数据集大多由动画类人模型构成，这些方法难以泛化到类人角色之外。其他方法 [4, 19, 21, 29, 38, 40, 46, 49, 55, 56, 61, 64, 67, 75, 78, 79, 83, 85]则通过从视频生成模型进行4D重建或蒸馏来避免监督学习，但由于高维4D运动难以优化，且引导信号中存在噪声，这些方法通常只能生成幅度较小且不真实的运动

。我们的框架通过不依赖特定类别的运动学结构或大规模4D数据集来解决这些局限，并为任意物体生成逼真的4D运动。

场景级4D生成。场景级4D生成超越了以物体为中心的设定，带来了更多复杂性和更大的挑战。它不仅必须生成合理的物体级运动，还必须在多个相互作用的物体之间保持运动一致性。因此，现有方法通常通过将问题限制在特定类别（例如，人-物体交互 [25, 37, 68, 81]）、施加物理约束 [9, 39, 41, 87], 或以符号结构作为条件 [3, 57]来简化问题。一些方法尝试通过重建视频模型生成的视频来生成4D场景 [10, 35, 66, 70]，但所得表示仍大体上是2.5D，无法支持完整的 360° 视图合成。我们首次解决了这一具有挑战性的设定：生成物体的场景级4D运动，而不依赖任何类别特定归纳偏置。

4D表示。4D生成流程中的一个关键环节是选择底层的4D表示。早期工作使用高维变形场来表示4D场景 [19, 50, 54, 69]。这类方法在输入稠密的重建目标上表现良好，但不适用于监督信号含噪的生成任务。近期工作探索了在空间域降低4D表示的维度 [21, 22, 66, 72]。我们的分层4D表示进一步发展了这一思路，在空间域和时间域中同时引入低维表示与层次结构，并将其作为4D生成框架中的骨干表示。

3. 方法

给定一个包含多个动态物体的3D场景，这些物体由各自的静态3D快照表示；同时给定一段文本提示，描述场景应如何随时间变化（例如，一名男子面对一盏台灯，文本提示为“男人用手把台灯的头往下压”），我们的目标是生成一系列时间变形来驱动物体，使最终生成的3D动画与文本提示相符。

图 2展示了我们方法的概览。我们通过迭代优化一种4D场景运动表示，并使用从视频生成模型中蒸馏出的引导信号。在以下章节中，我们将详细介绍该框架的三个主要组成部分：从现代基于整流流的视频生成模型中进行蒸馏的策略（第 3.2节）、一种稳健且通用的4D场景运动表示（第3.3节），以及用于确保优化稳定的正则化项（第 3.4节）。

3.1. 预备知识：分数蒸馏采样

分数蒸馏采样（SDS）方法 [53] 最初用于从图像扩散模型中蒸馏 3D 资产 [24]。每次迭代中，都会从由参数 θ 化的 3D 资产渲染出一张图像 $\mathbf{z}$ ϵ 随后加入高斯噪声，从而生成噪声图像 $\mathbf{z}_\tau$，其中噪声水平 τ 从 (0,1)中均匀采样。随后将噪声图像 $\mathbf{z}_\tau$ 输入图像扩散模型，由该模型预测噪声 $\hat{\epsilon}$。SDS 的更新方式 θ 如下，其梯度为:

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\theta; \mathbf{z}, \mathbf{y}) = \mathbb{E}_{\tau,\epsilon}\left[w(\tau) \left(\hat{\epsilon}\left(\mathbf{z}_\tau; \tau, \mathbf{y}\right) - \epsilon \right) \frac{\partial \mathbf{z}}{\partial \theta} \right], \tag{1}$$

其中， $w(\tau)$ 是一个加权函数。

将这一思路扩展到四维生成，遵循相同的原则：在每次迭代中，先从四维资产渲染出视频，再将其与噪声混合，然后输入扩散模型，由该模型提供梯度以更新 4D 表示。

3.2. 从整流流模型中蒸馏

上述四维分数蒸馏采样算法在概念上很简单，但要将其应用于现代视频生成模型的蒸馏并非易事。主要障碍在于，原始分数蒸馏采样目标所使用的扩散架构，与现代视频生成模型（例如我们论文中使用的 Wan2.2 [65]）所采用的基于 Rectified Flow（RF）的模型架构之间存在差异。

为缩小这一架构差异，我们为整流流模型推导出一种新颖的分数蒸馏采样目标。与为扩散模型推导分数蒸馏采样梯度的过程类似 [53], ，我们将优化目标与模型的训练损失对齐，并将整流流模型的分数蒸馏采样更新规则表示为:

$$\nabla_\theta \mathcal{L}_{\text{RFSDS}}(\theta; z, \mathbf{y}) = \mathbb{E}_{\tau,\epsilon}\left[w(\tau) \left(\hat{v}\left(\mathbf{z}_\tau; \tau, \mathbf{y}\right) - \epsilon + \mathbf{z} \right) \frac{\partial \mathbf{z}}{\partial \theta} \right], \tag{2}$$

其中， τ 是从 (0,1) 均匀采样得到的噪声水平，$w(\tau)$ 是训练调度中的对应权重，ϵ是添加的噪声，$\mathbf{z}_\tau = (1-\tau)\mathbf{z} + \tau\epsilon$表示噪声视频，而 $\hat{v}(\mathbf{z}_\tau; \tau, \mathbf{y})$ 是预测速度。

领域特定的噪声采样策略对于实现我们的目标——优化场景形变——至关重要。我们观察到，变形往往会在较高的噪声级别 τ 下生成，因为只有添加大量噪声时才会发生显著变化。基于这一观察以及$w(\tau)$ 的特性，我们不再均匀采样 τ，而是根据概率密度函数 $\hat{w}(\tau) = \frac{1}{w(\tau)} / \int_{-\infty}^{\infty} \frac{1}{w(\tau)} \, \mathrm{d}\tau$ $w(\tau)$ 进行采样，该函数是 $w(\tau)$ 的归一化形式。

通过对采样策略进行这一修改，加权后的

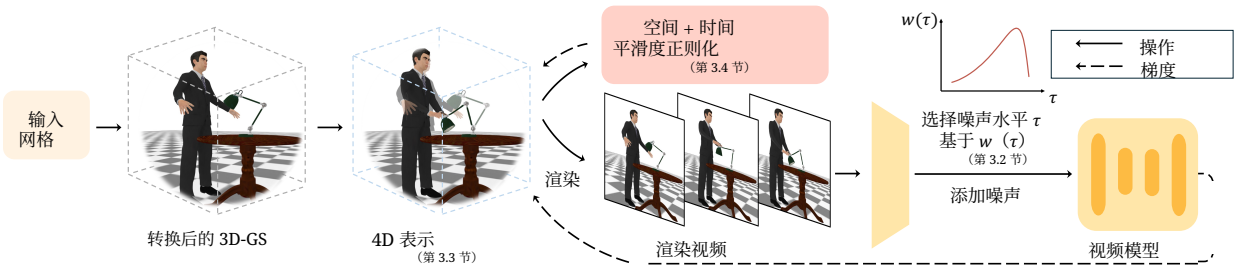

图 2. **概述**。 对于给定场景的输入网格，我们首先将其转换为 3D-GS 表示，以便进行平滑的梯度计算。然后，使用转换后的 3D-GS 模型初始化 4D 表示（第 3.3 节）。在每次迭代中，我们通过采样相机姿态、渲染相应的视频，并将其输入视频生成模型以获得优化梯度，来迭代优化此 4D 表示（第 3.2 节）。此外，我们还计算正则化项（第 3.4 节），以在优化过程中确保时空平滑性。

RFSDS 更新规则变为：

$$\nabla_\theta \mathcal{L}_{\text{W-RFSDS}}(\theta; z, \mathbf{y}) = \mathbb{E}_{\tau\sim\hat{w}(\tau),\epsilon}\left[\left(\hat{v}\left(\mathbf{z}_\tau; \tau, \mathbf{y}\right) - \epsilon + \mathbf{z}\right)\frac{\partial \mathbf{z}}{\partial \theta}\right], \tag{3}$$

其中，公式（16）中 RFSDS 梯度的权重项被消去，以确保梯度期望的不变性。从经验上看，这会产生更真实的生成运动，如第 4.3 节所示。

实际上，这种噪声采样策略通过优化过程中的退火噪声调度 [26, 62]来实现。在整个 i 次 I 迭代中的每个优化步骤 τ，我们将其设为固定的噪声水平 τ_i，该水平通过求解以下方程得到：

$$h(\tau_i) = 1 - \frac{i}{I+1}, \tag{4}$$

其中 $h(\tau) = \int_{-\infty}^{\tau} \hat{w}(t)\,\mathrm{d}t$ 是 $\hat{w}(\tau)$ 的累积分布函数（CDF）。这构建了一种退火调度，使 τ 在训练过程中逐渐减小，从而使粗略运动在早期形成，并允许在后续迭代中进一步优化精细形变。

3.3. 分层4D表示

现有的大多数 4D 表示在使用上述 W-RFSDS 目标进行优化时都非常不稳定。因此，我们引入了一种分层 4D表示，利用形变在空间域和时间域中的自然局部性，以稳定优化过程。

我们的表示由两个部分组成：规范形状的几何表示，以及一种 4D 运动表示，它会使规范几何在不同帧中发生形变。我们的 4D 表示中的规范形状由 3D-GS [31] 表示。具体而言，给定 N 网格输入，我们通过直接在从网格渲染得到的多视图图像上进行优化，将其转换为 3D-GS 模型$\mathcal{S} = \{\mathcal{G}_i\}_{i=1}^{N}$，其中每个 $\mathcal{G}_i$ 表示一个转换后的 3D-GS 模型。

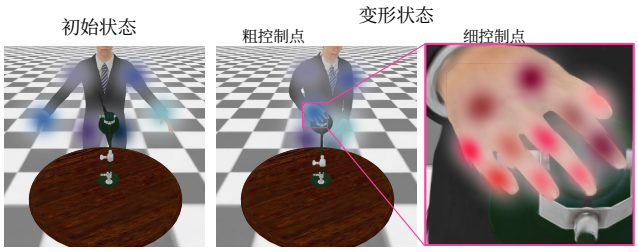

图 3。**分层控制点表示示意图**。我们使用空间层次结构表示形变。粗控制点捕捉大尺度形变，而细控制点则细化局部细节。

在时间 t，通过一组变形场对规范 3D 几何进行形变，以表示 3D-GS 场景 $\mathcal{S}$ 的 4D 运动。时间 t 的变形场集合记为 $\mathcal{D}^t = \{\mathcal{T}_i^t\}_{i=1}^{N}$，其中 $\mathcal{T}_i^t$表示物体 i 在时间 t 的形变。

4D场景运动 $\mathcal{D}$ 采用一种新颖的表示方式，该方式在空间域和时间域中均引入了层次结构，具体如下所述。

带控制点的空间层级。 形变场 $\mathcal{T}_i^t$具有很高的空间维度，因此我们通过分层的基于控制点的表示来降低该表示的维度。

受 SC-GS [27], 启发，我们用粗层和细层级的控制点来表示 $\mathcal{T}_i^t$——一组稀疏且定位于空间中的斑块，用于控制局部空间区域内的变形。粗层的控制点大致决定物体将如何形变，而细层级则为形变增添更多细节。

具体而言，每个控制点由均值 $\boldsymbol{p}$和协方差矩阵 $\boldsymbol{\Sigma}$ 定义，二者共同决定其影响半径。此外，每个控制点都会在 $SE(3)$ 中维持一组形变序列（$\mathbf{R}^t$,$\mathbf{T}^t$）。高斯的形变通过线性混合蒙皮，对来自相邻控制点的变换进行混合而得到。对于一个高斯（$\boldsymbol{\mu}, \boldsymbol{q}, \boldsymbol{S}, \mathcal{C}, o$） $\in \mathcal{G}_i$，我们将其最近邻控制点记为 $\mathcal{N}$。形变后的

时刻 t 的高斯随后计算如下：

$$\mu^t = \sum_{k\in\mathcal{N}} \beta_k \left(R_k^t(\mu - \mathbf{p}_k) + \mathbf{p}_k + T_k^t\right), \tag{5}$$

$$\mathbf{q}^t = (\sum_{k\in\mathcal{N}} \beta_k r_k^t) \otimes \mathbf{q}, \tag{6}$$

其中，$r_k^t \in \mathbb{R}^4$ 是控制点 k上的旋转四元数表示，$\otimes$ 是四元数的乘积。此外，公式中的 β_k 表示控制点 k的混合权重，其计算方式如下：

$$\beta_k = \frac{\hat{\beta}_k}{\sum_{l\in\mathcal{N}} \hat{\beta}_l}, \ \hat{\beta}_k = \exp\left(-\frac{1}{2}\left[(\mu - p_k)\,\Sigma_k^{-1}\,(\mu - p_k)^T\right]\right). \tag{7}$$

我们按照公式（4）定义的噪声调度，以由粗到细的方式优化双层控制点集。当 τ 在优化过程中较大时，可以生成显著的运动；然而，在此类噪声级别下产生的 SDS 梯度通常带有较大噪声。相反，当 τ 退火至较低值时，梯度会变得更加稳定，但生成显著变形的能力也会减弱。为适应这一优化过程的内在特性，当 τ 较大时，我们仅在早期迭代中优化粗层控制点；待 τ 变小时，再引入细控制点，以补充其残余变形：

$$\mu_{\text{final}}^t = \Delta\mu^t + \mu_t, \tag{8}$$

$$\mathbf{q}_{\text{final}}^t = \Delta\mathbf{q}^t \otimes \mathbf{q}^t \tag{9}$$

其中 $\Delta\boldsymbol{\mu}^t$ 和 $\Delta\mathbf{q}^t$ 表示来自控制点细层的残余变形，其计算方式与公式（5）和公式（6）相同。

训练完成后，可以将通过高斯学习到的形变直接迁移到网格上。具体而言，我们将高斯均值替换为顶点位置，并使用公式（5）对网格顶点进行形变。

基于芬威克树的时间层级。我们进一步发现，如果将帧 t 的 (R^t, T^t) 独立于其他帧进行建模，后续帧的变形就很难学习。这是因为所有变形最初都初始化为零向量，并且第一帧的参数保持冻结，导致后续帧的变形出现显著偏差。

为缓解这一问题，我们使用芬威克树表示每个控制点 $(R^t,\ T^t)$ 的形变序列。芬威克树是一种源自理论算法设计的层次数据结构 [18]。如图 4所示，对于每个控制点 k，我们维护节点 $\mathcal{F}_k = \{(r_k^{[j]},\ T_k^{[j]})\}_{j=1}^T$，每个节点都编码特定帧范围内的累积变形。这种基于范围的分解使不同帧上的变形能够共享参数

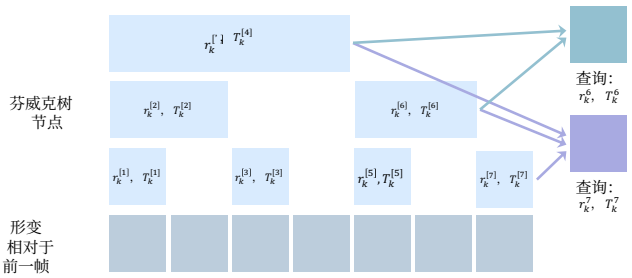

图 4. **芬威克树表示示意图**。每个节点存储一个时间范围内的累积变形，使相邻帧能够共享参数，并自然地保证时间一致性。例如，$(r_k^{[6]}, T_k^{[6]})$ 编码了帧 5–6 的累积变形。随后，对帧 6 和 7 的查询会如图所示，通过一小组相互重叠的节点组合出各自的变形。

通过重叠区间大幅提升时间一致性，并支持长时域运动的学习。

通过组合所有相关节点，可得到帧 t 处的最终形变：

$$T_k^t = \sum_{j\in\text{BIT}(t)} T_k^{[j]}, \tag{10}$$

$$r_k^t = \text{norm}(\sum_{j\in\text{BIT}(t)} r_k^{[j]}), \tag{11}$$

其中，BIT(t) 表示通过芬威克树查询操作返回的活动节点集合，而 norm(·) 确保求和结果构成有效的四元数。

3.4. 正则化

我们引入两个正则化项，以进一步稳定优化过程：时间正则化损失用于确保时间上的平滑性，空间正则化损失则用于增强局部空间一致性。

时间正则化。 在渲染 RGB 视频以计算 SDS 梯度时，我们还会从相同视点渲染三维流场图视频$\mathbf{F}$ ，用于时间正则化。为生成帧 t 处的流场图，我们将 3D-GS 渲染方程中高斯的颜色属性替换为 $\boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^{t+1}$ ，其中 $\boldsymbol{\mu}_i^t$ 表示时间 t 处高斯 i 的均值。得到 $\mathbf{F}$ 后，时间正则化损失定义如下：

$$\mathcal{L}_{\text{temp}} = \sum_t \sum_{\mathbf{p}} ||F_{\mathbf{p}}^t||_2^2, \tag{12}$$

其中，内层求和遍历所有像素$\mathbf{p}$，而$\mathbf{F}_{\mathbf{p}}^t$ 表示时间 t、像素$\mathbf{p}$处渲染得到的三维流场。

空间正则化。为确保空间上均匀的正则化，我们在每个物体 i的表面附近生成均匀分布的点云，使用学习到的运动对其进行变形，并在所得的变形点云序列上计算尽可能刚性（ARAP）

损失 [59] 。具体而言，我们首先根据物体 i 的网格计算有符号距离场（SDF） $\phi_i(\mathbf{x})$。然后提取表面附近的体素中心，得到 $\mathcal{S}_i = \{\mathbf{x} \mid |\phi_i(\mathbf{x})| \leq \tau, \mathbf{x} \in V_s\}$，其中 V_s 是网格上体素大小为 s 的体素中心集合， τ 是预定义的阈值。在每次迭代中，对于每个 $\mathbf{x} \in \mathcal{S}_i$ 和时间戳 t，我们使用公式（5）计算其变形后的位置 $\mathbf{x}^t$（将 μ 替换为 $\mathbf{x}$），从而生成变形点集 $\mathcal{S}_i^t = \{\mathbf{x}^t \mid \mathbf{x} \in \mathcal{S}_i\}$。随后计算 ARAP损失：

$$\mathcal{L}_{\text{ARAP}} = \sum_{i,t,\mathbf{x}\in\mathcal{S}_i,\mathbf{y}\in\mathcal{N}_\mathbf{x}} ||\mathbf{x} - \mathbf{y} - \hat{R}_\mathbf{x}(\mathbf{x}^t - \mathbf{y}^t)||_2^2, \quad (13)$$

其中， $\mathcal{N}_\mathbf{x}$ 表示 $\mathcal{S}_i$中$\mathbf{x}$ˆ的 10 个最近邻组成的集合， $R_\mathbf{x}$ 是 $\mathbf{x}$处估计的局部旋转矩阵。

4. 实验

我们在包含多个交互物体的多种动态场景上评估所提出的方法。我们将该方法与多种最先进的基线方法进行比较，每种方法代表一类不同的方法。

4.1. 4D场景运动生成

我们将该方法与最先进的网格动画方法，以及来自相机控制视频模型的 4D重建进行比较。具体而言，我们将该方法与四种基线方法进行比较：Animate3D [30], AnimateAnyMesh [73], MotionDreamer [64], 和 TrajectoryCrafter [84]。Animate3D 使用多视角视频扩散模型生成多视角视频，然后对其进行 4D重建。AnimateAnyMesh 使用预训练的 Rectified Flow 模型直接预测网格变形。MotionDreamer 首先根据文本提示和给定网格的渲染结果生成视频，然后通过与生成的视频进行扩散特征匹配来为网格添加动画。我们展示了使用 Wan 2.2 重新实现所得的结果，并在补充材料中提供了使用 DynamiCrafter 得到的结果 [76], 它已用于原始流程中。TrajectoryCrafter 是一种视频生成模型，可重新定向单目视频的相机轨迹。我们首先使用 Wan 2.2 生成视频，然后用 TrajectoryCrafter 生成相应的多视角视频，最后对采样得到的视频进行 4D重建。

我们选取涵盖多种物体类别的六个场景进行比较：“一名男子抚摸一只狗”、“一只猫踩在坐垫上”、“一只海狮用鼻子推球”、“一个方块落在蹦床上”、“两名男子握手”以及“一台机器人捡起一个方块”。此外，我们还在补充材料中加入了我们的方法与基线方法在**单物体网格动画**方面的比较。

定性比较。 部分定性结果如图5所示；完整结果集请参阅补充材料

表1. **与基线方法的定量比较**。 我们针对六段场景动画开展了用户研究，以评估性能。此外，我们还报告了使用 VideoPhy-2 [5] 计算得出的语义一致性（SA）和物理常识（PC）指标。

	用户研究		VideoPhy-2	
	对齐度 ↑	真实感 ↑	SA ↑	PC ↑
Animate3D	0.34%	0.51%	3.83	3.42
AnimateAnyMesh	1.01%	0.51%	3.5	**4.5**
MotionDreamer（DC）	0.51%	0.84%	3.42	4.08
MotionDreamer（Wan）	0.84%	0.34%	3.5	3.83
TrajectoryCrafter	9.60%	10.44%	<u>4.17</u>	3.83
CHORD（我们的方法）	**87.71%**	**87.37%**	**4.33**	<u>4.25</u>

。与现有方法相比，我们的方法具有更强的提示词对齐能力，生成的运动也更加自然。Animate3D 和 AnimateAnyMesh 未能生成与给定提示相符的结果，因为它们并未针对包含多个物体的 4D 数据进行充分训练。MotionDreamer 在拟合网格时，由于扩散特征匹配存在误差，产生了严重的伪影。尽管通过 TrajectoryCrafter 对视频进行采样后再进行 4D重建，所得运动能够遵循提示，但由于不同相机轨迹下生成的视频之间存在差异，结果出现了严重的时间不一致和不自然的动态效果。这凸显了在我们的方法中对视频模型进行蒸馏的必要性。

定量比较。我们对 99 位参与者开展了用户研究，以比较我们的方法与基线方法的质量。此外，我们使用 VideoPhy-2 [5] 从两个方面自动评估渲染视频：语义一致性（SA）和物理常识（PC）。如表1所示，我们的方法在 SA 上得分最高，在 PC 上排名第二。值得注意的是，AnimateAnyMesh 的物理常识（PC）得分最高，这是因为它常见的一种失败模式是物体保持静止——这符合物理常识，却未能遵循给定的提示。

4.2. 扩展与应用

除了生成多物体四维运动外，我们的框架还可自然地支持多种扩展和下游应用。

长时域运动生成。将生成形变的最后一帧用作后续生成过程的输入状态，即可扩展我们的方法，生成更长的运动序列。图1展示了一个由四个动作组成的运动序列示例。

真实世界物体动画。由于我们的方法提炼了一个经过大量真实世界视频数据训练的视频生成模型，因此具有很强的鲁棒性，可以用于为扫描的真实世界物体制作动画，而不必担心合成数据与真实世界数据之间的差异，如图6所示。

机器人操作。我们展示了，通过我们的方法生成的稠密物体

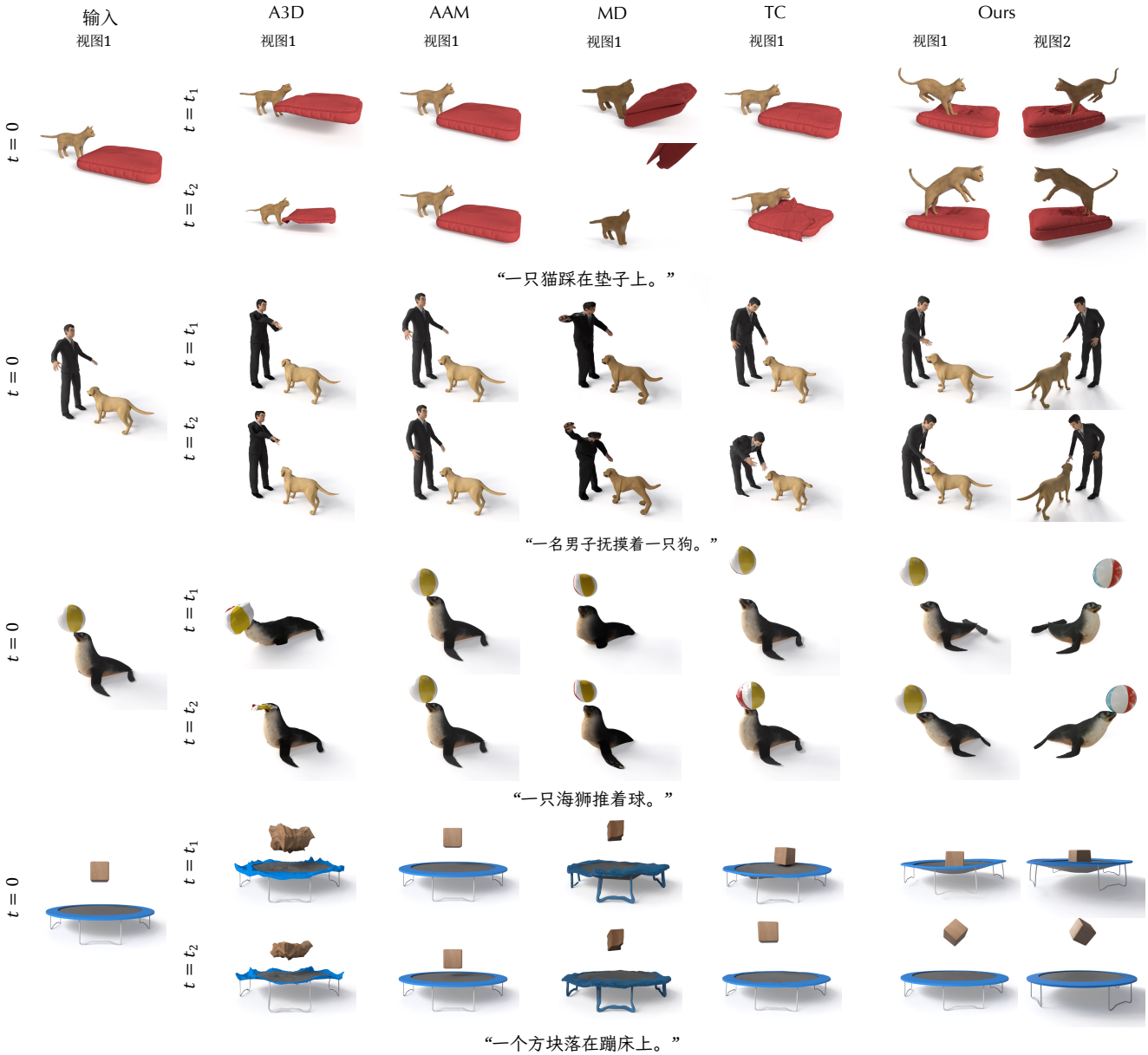

图5。 **定性比较**。 我们将我们的方法与几种网格动画方法进行了比较。与给定提示词的契合度相比，我们的方法生成的结果更好，并呈现出更自然的运动。在图中，A3D指的是Animate3D [30], AAM表示AnimateAnyMesh [73], MD代表MotionDreamer [64], , TC对应于TrajectoryCrafter生成的视频中的4D重建结果 [84]。如需查看更多比较结果和完整的动画结果，请参阅我们的补充材料网站。



“一名男子合上笔记本电脑的盖子。”

图6。 **真实世界物体动画结果**。

流可用于引导对刚性、关节式和可变形物体的操作，如图7所示。我们首先使用现成的抓取规划器 [17]用于在相关物体上提出抓取方案。随后，机器人要么抓取该物体，要么移动到一个姿态以推动物体，该姿态与所提议的抓取位置存在偏移。在刚性附着前向模型的约束下，末端执行器的相对变换也会作用于物体上的初始点；运动规划器 [32] 求解一系列末端执行器姿态，以最小化由变换后的点与稠密流对齐、可达性以及位姿平滑度代价构成的目标函数。

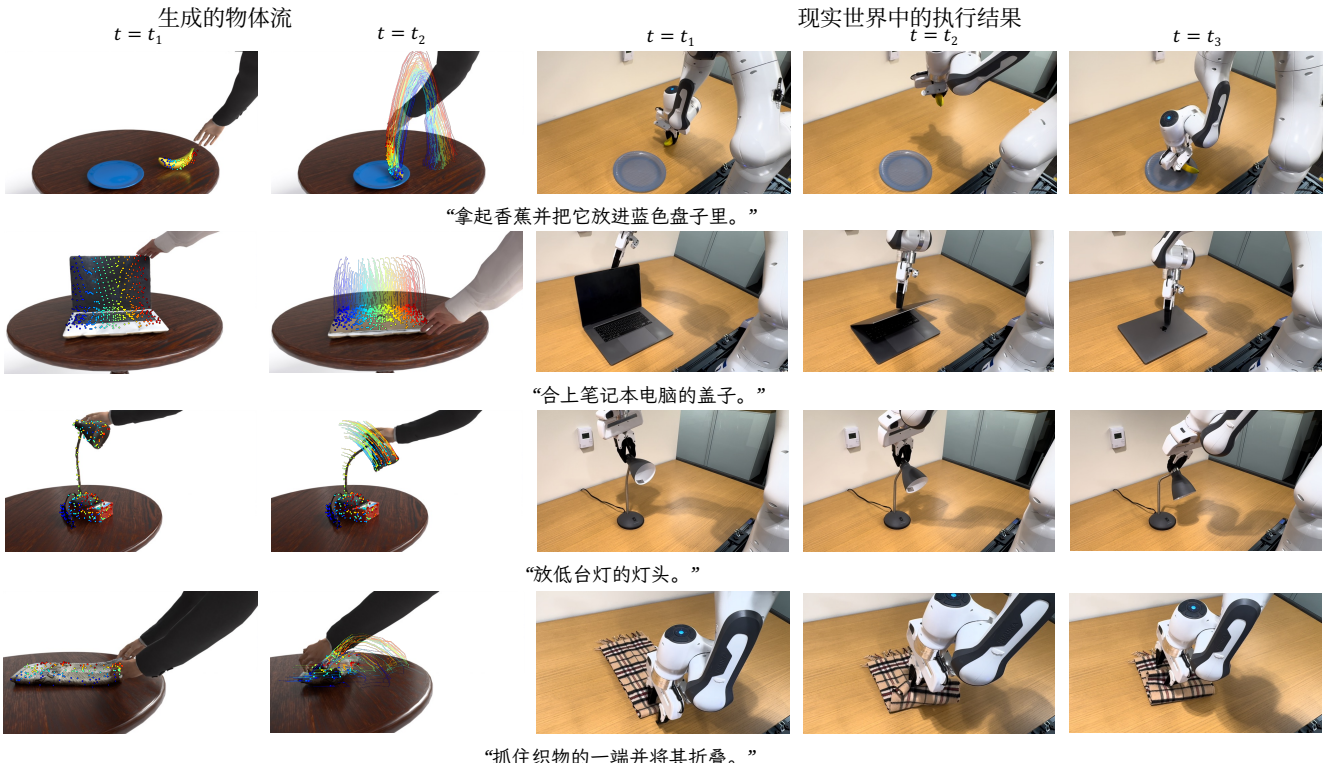

图7. **由我们生成的稠密物体流引导的机器人操作**。 根据我们生成的稠密物体流，机器人以符合流动方向的方式抓取或推动目标物体。这使得机器人能够有效操作刚体（第一行）、关节物体（第二行）和可变形物体（第三、四行）。



图8. **噪声级采样策略的消融实验**。 移除我们的噪声级采样策略会导致不自然的运动，例如笔记本电脑看起来像是漂浮起来了。



图9. **对4D表示中各组件的消融实验**。 移除芬威克树会导致后续帧中出现严重伪影；移除细控制点会使细节变形无法实现；移除粗控制点则会造成失真。

4.3. 消融研究

噪声级采样策略我们将噪声级采样策略（第3.2节）与带权重的均匀噪声采样进行比较。如图8所示，由于均匀采样对注入运动的噪声级别覆盖不足，因此会产生不切实际的结果。在这种情况下，笔记本电脑看起来像是悬浮在桌面上方。

4D 表示。 我们研究了 4D 表示的两个关键组成部分：用于建模形变序列的芬威克树，以及分层控制点结构。结果如图 9所示。移除芬威克树会导致明显的伪影，因为如果分别学习每个形变，后续帧将变得极难学习。

移除细控制点层会导致模型无法生成精细运动（例如抓握动作）。反之，从一开始就使用细层也会引入伪影，因为如果没有初始的粗阶段，早期迭代中注入的噪声就无法得到有效平滑。

正则化。 我们证明了正则化损失是必不可少的。移除这些损失会导致时间闪烁（例如，移除时间正则化后，尾巴会突然出现）和视觉伪影（移除空间正则化后），如图 10所示。



图10。**正则化损失的消融实验**。移除时间正则化会导致闪烁，而移除空间正则化则会导致失真。

5. 结论

我们提出了一种稳健、可扩展且用途广泛的方法，仅以3D形状为输入即可生成场景级 4D物体运动。我们的流程能够有效处理各种自然现象，并借助视频生成模型的引导，为可扩展的 4D 生成开辟了新的可能性。正如我们在机器人操作中所展示的，该方法还支持下游应用。

参考文献

[1] Sketchfab。https://sketchfab.com。访问日期：2025-11-18。13[2] Noam Aigerman、Kunal Gupta、Vladimir G. Kim、Siddhartha Chaudhuri、Jun Saito 和 Thibault Groueix。神经雅可比场：学习任意网格的内在映射。*ACM* 图形学汇刊（*SIGGRAPH 2022*），2022年。13[3]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。载于欧洲计算机视觉会议，第53–72页。Springer，2024年。3[4] 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进行文本到4D生成。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第7996–8006页，2024年。2[5] Hritik Bansal、Clark Peng、Yonatan Bitton、Roman Goldenberg、Aditya Grover 和 Kai-Wei Chang。VideoPhy-2：一项具有挑战性的、以动作为中心的视频生成物理常识评估。*arXiv*预印本 *arXiv:2503.06800*，2025年。6[6] Volker Blanz 和 Thomas Vetter。基于拟合3D可变形模型的人脸识别。*IEEE* 模式分析与机器智能汇刊，25(9):1063–1074，2003年。2[7] BlenderKit。BlenderKit在线素材库。https://www.blenderkit.com。访问日期：2025-11-18。13[8] James Booth、Anastasios Roussos、Stefanos Zafeiriou、Allan Ponniah 和 David Dunaway。基于10,000张人脸学习得到的3D可变形模型。载于*IEEE*会议论文集计算机视觉与模式识别，页码 5543 5552，2016 2 – 。[9] Boyuan Chen、Hanxiao Jiang、Shaowei Liu、Saurabh Gupta、Yunzhu Li、Hao Zhao 和 Shenlong Wang。Physgen3d：从单张图像打造一个微型互动世界。载于计算机视觉与模式识别会议论文集，页码 6178–6189，2025。3[10] Wen-Hsuan Chu、Lei Ke 和 Katerina Fragkiadaki。Dream-scene4d：从单目视频生成动态多物体场景。神经信息处理系统进展，37：96181–96206，2024。3[11] Blender Online Community。Blender——一个三维建模与渲染软件包。Blender Foundation、Stichting Blender Foundation，阿姆斯特丹，2018。13[12] 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 10m+ 3d 个物体的宇宙。神经信息处理系统进展，36：35799–35813，2023。2[13] Matt Deitke、Dustin Schwenk、Jordi Salvador、Luca Weihs、Oscar Michel、Eli VanderBilt、Ludwig Schmidt、Kiana Ehsani、Aniruddha Kembhavi 和 Ali Farhadi。Objaverse：一个标注的三维物体宇宙。载于IEEE/CVF计算机视觉与模式识别会议论文集，页码 13142–13153，2023。2[14] Yufan Deng、Yuhao Zhang、Chen Geng、Shangzhe Wu 和 Jiajun Wu。Anymate：用于学习三维物体绑定的数据集与基线方法。载于计算机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[15] E Knuth Donald 等。计算机程序设计艺术。排序与查找，3(426-458):4，1999。2[16] Patrick Esser、Sumith Kulal、Andreas Blattmann、Rahim Entezari、Jonas M¨uller、Harry Saini、Yam Levi、Dominik Lorenz、Axel Sauer、Frederic Boesel 等。扩展用于高分辨率图像合成的整流流变换器。载于第 41 届国际机器学习会议论文集，页码 12606–12633，2024。13[17] Hao-Shu Fang、Chenxi Wang、Hongjie Fang、Minghao Gou、Jirong Liu、Hengxu Yan、Wenhai Liu、Yichen Xie 和 Cewu Lu。Anygrasp：在空间域和时间域中实现稳健高效的抓取感知。IEEE机器人学汇刊（T-RO），2023。7[18] Peter M Fenwick。用于累积频率表的新型数据结构。Software: Practice and experience，24(3)：327–336，1994。5[19] Quankai Gao、Qiangeng Xu、Zhe Cao、Ben Mildenhall、Wen-chao Ma、Le Chen、Danhang Tang 和 Ulrich Neumann。Gaussianflow：泼溅高斯动态以实现四维内容创作。arXiv 预印本 arXiv:2403.12365，2024。2，3[20] Chen Geng、Sida Peng、Zhen Xu、Hujun Bao 和 Xiaowei Zhou。数分钟内学习动态人类的神经体积表示。载于IEEE/CVF计算机视觉与模式识别会议论文集，页码 8759–8770，2023。2[21] Chen Geng、Yunzhi Zhang、Shangzhe Wu 和 Jiajun Wu。玫瑰的诞生与凋亡。载于计算机

视觉与模式识别会议论文集，页码 26102–26113，2025。2，3[22] Guangzhao He、Chen Geng、Shangzhe Wu 和 Jiajun Wu。类别无关神经对象绑定。载于计算机视觉与模式识别会议论文集，页码 22078–22088，2025。2，3[23] Jonathan Ho 和 Tim Salimans。无分类器扩散引导。arXiv 预印本 arXiv:2207.12598，2022。13[24] Jonathan Ho、Ajay Jain 和 Pieter Abbeel。去噪扩散概率模型。载于神经信息处理系统进展（NeurIPS），2020。3[25] Di Huang、Xiaopeng Ji、Xingyi He、Jiaming Sun、Tong He、Qing Shuai、Wanli Ouyang 和 Xiaowei Zhou。从单目视频重建手持物体。载于 SIGGRAPH Asia 2022 会议论文集，页码 1–9，2022。3[26] Yukun Huang、Jianan Wang、Yukai Shi、Boshi Tang、Xian-biao Qi 和 Lei Zhang。Dreamtime：一种改进的扩散引导三维生成优化策略。载于第十二届国际学习表征会议（ICLR），2024。4[27] Yi-Hua Huang、Yang-Tian Sun、Ziyi Yang、Xiaoyang Lyu、Yan-Pei Cao 和 Xiaojuan Qi。SC-GS：用于可编辑动态场景的稀疏控制高斯泼溅。载于 IEEE/CVF计算机视觉与模式识别会议论文集，页码 4220–4230，2024。4[28] Zehuan Huang、Haoran Feng、Yangtian Sun、Yuanchen Guo、Yanpei Cao 和 Lu Sheng。Animax：利用联合视频姿态扩散模型为三维无生命物体制作动画。arXiv 预印本 arXiv:2506.19851，2025。2[29] Yanqin Jiang、Li Zhang、Jin Gao、Weimin Hu 和 Yao Yao。Consistent4d：从单目视频生成一致的 360 {\deg} 动态物体。arXiv 预印本 arXiv:2311.02848，2023。2[30] Yanqin Jiang、Chaohui Yu、Chenjie Cao、Fan Wang、Weim-ing Hu 和 Jin Gao。Animate3d：利用多视图视频扩散为任意三维模型制作动画。神经信息处理系统进展，37：125879–125906，2024。2，6，7，13，14，15，18[31] Bernhard Kerbl、Georgios Kopanas、Thomas Leimk¨uhler 和 George Drettakis。用于实时辐射场渲染的三维高斯泼溅。ACM 图形学汇刊，42 1，2023 4 139 4，13 ( ): – 。[32] Chung Min Kim*、Brent Yi*、Hongsuk Choi、Yi Ma、Ken Goldberg 和 Angjoo Kanazawa。Pyroki：用于机器人运动学优化的模块化工具包。载于 2025 年 IEEE/RSJ 国际智能机器人与系统会议（IROS），2025。7[33] Diederik P. Kingma 和 Max Welling。自动编码变分贝叶斯。载于国际学习表征会议（ICLR）论文集，2014。16[34] Sumith Kulal、Jiayuan Mao、Alex Aiken 和 Jiajun Wu。通过运动程序实现分层运动理解。载于 IEEE/CVF计算机视觉与模式识别会议论文集，页码 6568–6576，2021。2[35] Jiahui Lei、Yijia Weng、Adam W Harley、Leonidas Guibas 和 Kostas Daniilidis。Mosca：通过四维运动支架从随拍视频进行动态高斯融合。载于论文集

视觉与模式识别会议，第 6165–6177 页，2025 年。 3[36] Chengshu Li、Ruohan Zhang、Josiah Wong、Cem Gokmen、Sanjana Srivastava、Roberto Mart´ın-Mart´ın、Chen Wang、Gabrael Levine、Michael Lingelbach、Jiankai Sun 等。Behavior-1k：面向具身人工智能的基准，涵盖 1,000 项日常活动与真实感仿真。载于机器人学习会议，第 80–93 页。PMLR，2023 年。 2[37] Hongjie Li、Hong-Xing Yu、Jiaman Li 和 Jiajun Wu。Zerohsi：通过视频生成实现零样本四维人景交互。arXiv 预印本 arXiv:2412.18600，2024 年。3[38] Zhiqi Li 、 Yiming Chen 和 Peidong Liu。Dreammesh4d：采用稀疏控制的高斯-网格混合表示进行视频到四维生成。神经信息处理系统进展，37：21377–21400，2024 年。 2[39] Zizhang Li、Hong-Xing Yu、Wei Liu、Yin Yang、Charles Hermann、Gordon Wetzstein 和 Jiajun Wu。Wonderplay：从单张图像和动作生成动态 3D场景。arXiv 预印本 arXiv:2505.18151，2025 年。 3[40]Hanwen Liang、Yuyang Yin、Dejia Xu、Hanxue Liang、Zhangyang Wang、Konstantinos N Plataniotis、Yao Zhao 和 Yunchao Wei。Diffusion4d：通过视频扩散模型快速生成具有时空一致性的四维内容。arXiv 预印本 arXiv:2405.16645，2024 年。 2[41] Yuchen Lin、Chenguo Lin、Jianjin Xu 和 Yadong Mu。Omniphysgs：用于通用基于物理的动态生成的三维本构高斯。arXiv 预印本 arXiv:2501.18982，2025 年。 3[42] Lingjie Liu、Marc Habermann、Viktor Rudnev、Kripasindhu Sarkar、Jiatao Gu 和 Christian Theobalt。Neural Actor：带姿态控制的人体角色神经自由视角合成。ACM 图形学汇刊（TOG），40：16，2021 年。 6 1 2 ( ): – 。[43] Ruoshi Liu、Alper Canberk、Shuran Song 和 Carl Vondrick。可微机器人渲染。arXiv 预印本 arXiv:2410.13851，2024 年。 2[44] Xingchao Liu、Chengyue Gong 和 Qiang Liu。Flow straight and fast：使用 Rectified Flow 学习生成和迁移数据。载于第十一届国际学习表征会议（ICLR），2023 年。 2, 13[45] Matthew Loper、Naureen Mahmood、Javier Romero、Gerard Pons-Moll 和 Michael J Black。SMPL：带蒙皮的多人线性模型。载于图形学奠基论文：拓展边界（第二卷），第 851–866 页。2023 年。2[46] Ziqiao Ma、Xuweiyi Chen、Shoubin Yu、Sai Bi、Kai Zhang、Chen Ziwen、Sihan Xu、Jianing Yang、Zexiang Xu、Kalyan Sunkavalli 等。4D-LRM：从任意视角到任意时刻的大型时空重建模型。arXiv 预印本 arXiv:2506.18890，2025 年。 2[47] Naureen Mahmood、Nima Ghorbani、Nikolaus F Troje、Gerard Pons-Moll 和 Michael J Black。AMASS：动作捕捉表面形状档案。载于 IEEE/CVF 国际计算机视觉会议论文集，第 5442–5451 页，2019 年。 2[48] Zhao Mandi、Yifan Hou、Dieter Fox、Yashraj Narang、Ajay Mandlekar 和 Shuran Song。Dexmachina：面向双手灵巧操作的功能性重定向。arXiv 预印本 arXiv:2505.24853，2025 年。 2

[49] Linzhan Mou、Jiahui Lei、Chen Wang、Lingjie Liu 和 Kostas Daniilidis。Dimo：面向任意物体的多样化 3D 动作生成。载于 IEEE/CVF 国际计算机视觉会议论文集，第 14357 14368 页，2025 年 2 – 。[50] Keunhong Park、Utkarsh Sinha、Jonathan T Barron、Sofien Bouaziz、Dan B Goldman、Steven M Seitz 和 Ricardo Martin-Brualla。Nerfies：可变形神经辐射场。载于 IEEE/CVF 国际计算机视觉会议论文集，第 5865–5874 页，2021 年。 3[51]Georgios Pavlakos、Vasileios Choutas、Nima Ghorbani、Timo Bolkart、Ahmed AA Osman、Dimitrios Tzionas 和 Michael J Black。表现力人体捕捉：从单张图像获取 3D 手部、面部和身体。载于 IEEE/CVF 计算机视觉与模式识别会议论文集，第 10975–10985 页，2019 年。 2[52] Sida Peng、Chen Geng、Yuanqing Zhang、Yinghao Xu、Qian-qian Wang、Qing Shuai、Xiaowei Zhou 和 Hujun Bao。采用结构化潜在编码的隐式神经表示及其在人体建模中的应用。IEEE 模式分析与机器智能汇刊，45(8)：9895–9907，2023 年。 2[53] Ben Poole、Ajay Jain、Jonathan T. Barron 和 Ben Mildenhall。Dreamfusion：使用 2D 扩散进行文本转 3D。载于 ICLR。OpenReview.net，2023 年。 2, 3, 13[54] Albert Pumarola、Enric Corona、Gerard Pons-Moll 和 Francesc Moreno-Noguer。D-nerf：用于动态场景的神经辐射场。载于 IEEE/CVF 计算机视觉与模式识别会议论文集，第 10318–10327 页，2021 年。 3[55] Jiawei Ren、Liang Pan、Jiaxiang Tang、Chi Zhang、Ang Cao、Gang Zeng 和 Zi wei Liu。Dreamgaussian4d：生成式 4 17142，2023 年 2312 2D 高斯泼溅。arXiv 预印本 arXiv: . . [56] Jiawei Ren、Cheng Xie、Ashkan Mirzaei、Karsten Kreis、Ziwei 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 年。 2[57] Qing Shuai、Chen Geng、Qi Fang、Sida Peng、Wenhao Shen、Xiaowei Zhou 和 Hujun Bao。稀疏多视角视频中的人类交互新视角合成。载于 ACM SIGGRAPH 2022 10，2022 年 1 3 会议论文集，第 – 页。[58] Chaoyue Song、Xiu Li、Fan Yang、Zhongcong Xu、Jiacheng Wei、Fayao Liu、Jiashi Feng、Guosheng Lin 和 Jianfeng Zhang。Puppeteer：绑定并制作你的 3D 模型动画。arXiv 预印本 arXiv:2508.10898，2025 年。2[59] Olga Sorkine 和 Marc Alexa。尽可能刚性的曲面建模。载于几何处理研讨会，第 109–116 页，2007 年。 6[60] Shih-Yang Su、Frank Yu、Michael Zollh¨ofer 和 Helge Rhodin。A-nerf：用于学习人体形状、外观和姿态的可关节活动神经辐射场。神经信息处理系统进展，34：12278–12291，2021 年。 2[61] Qi Sun、Zhiyang Guo、Ziyu Wan、Jing Nathan Yan、Sheng-ming Yin、Wengang Zhou、Jing Liao 和 Houqiang Li。Eg4d：无需分数蒸馏的 4D 物体显式生成。arXiv 预印本 arXiv:2405.18132，2024 年。 2[62] Jiaxiang Tang、Jiawei Ren、Hang Zhou、Ziwei Liu 和 Gang Zeng。《Dreamgaussian：用于高效 3D 内容创建的生成式高斯泼溅》。发表于第十二届国际学习表征会议（ICLR），2024 年。4[63] 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 年。2[64] Lukas Uzolas、Elmar Eisemann 和 Petr Kellnhofer。《Motiondreamer：探索用于零样本 3D 网格动画的语义视频扩散特征》。发表于 2025 年国际 3D 视觉会议（3DV），第 893–904 页。IEEE，2025 年。2、6、7、13、14、15、18[65] Wan 团队、Ang Wang、Baole Ai、Bin Wen、Chaojie Mao、Chen-Wei Xie、Di Chen、Feiwu Yu、Haiming Zhao、Jianx- iao Yang、Jianyuan Zeng、Jiayu Wang、Jingfeng Zhang、Jingren Zhou、Jinkai Wang、Jixuan Chen、Kai Zhu、Kang Zhao、Keyu Yan、Lianghua Huang、Mengyang Feng、Ningyi Zhang、Pandeng Li、Pingyu Wu、Ruihang Chu、Ruili Feng、Shiwei Zhang、Siyang Sun、Tao Fang、Tianxing Wang、Tianyi Gui、Tingyu Weng、Tong Shen、Wei Lin、Wei Wang、Wei Wang、Wenmeng Zhou、Wente Wang、Wenting Shen、Wenyuan Yu、Xianzhong Shi、Xiaoming Huang、Xin Xu、Yan Kou、Yangyu Lv、Yifei Li、Yijing Liu、Yiming Wang、Yingya Zhang、Yitong Huang、Yong Li、You Wu、Yu Liu、Yulin Pan、Yun Zheng、Yuntao Hong、Yupeng Shi、Yutong Feng、Zeyinzi Jiang、Zhen Han、Zhi-Fan Wu 和 Ziyu Liu。《Wan：开放且先进的大规模视频生成模型》。arXiv 预印本 arXiv:2503.20314，2025 年。3、13、18[66] Qianqian Wang、Vickie Ye、Hang Gao、Weijia Zeng、Jake Austin、Zhengqi Li 和 Angjoo Kanazawa。《运动的形态：从单个视频进行 4D 重建》。发表于 IEEE/CVF 国际计算机视觉会议论文集，第 9660–9672 页，2025 年。3[67] Yikai Wang、Xinzhou Wang、Zilong Chen、Zhengyi Wang、Fuchun Sun 和 Jun Zhu。《Vidu4d：从单个生成视频进行高保真 4D 重建，采用动态高斯表面元》。神经信息处理系统进展，37：131316–131343，2024 年。2[68] Bowen Wen、Jonathan Tremblay、Valts Blukis、Stephen Tyree、Thomas M¨uller、Alex Evans、Dieter Fox、Jan Kautz 和 Stan Birchfield。《BundleSDF：对未知物体进行神经 6 自由度跟踪和 3D 重建》。发表于 IEEE/CVF 计算机视觉与模式识别会议论文集，第 606–617 页，2023 年。3[69] Guanjun Wu、Taoran Yi、Jiemin Fang、Lingxi Xie、Xiaopeng Zhang、Wei Wei、Wenyu Liu、Qi Tian 和 Xinggang Wang。《用于实时动态场景渲染的 4D 高斯泼溅》。发表于 IEEE/CVF 计算机视觉与模式识别会议论文集，第 20310–20320 页，2024 年。3[70] Rundi Wu、Ruiqi Gao、Ben Poole、Alex Trevithick、Changxi Zheng、Jonathan T Barron 和 Aleksander Holynski。《CAT4D：通过多视图视频扩散模型创造任意 4D 内容》。发表于计算机视觉与模式识别会议论文集，第 26057–26068 页，2025 年。3[71] Shangzhe Wu、Ruining Li、Tomas Jakab、Christian Rup- precht 和 Andrea Vedaldi。《MagicPony：在野外学习可关节运动的 3D 动物》。发表于

IEEE/CVF 计算机视觉与模式识别会议论文集，第 8792–8802 页，2023 年。2[72] Zijie Wu、Chaohui Yu、Yanqin Jiang、Chenjie Cao、Fan Wang 和 Xiang Bai。《Sc4d：稀疏控制的视频到四维生成与运动迁移》。发表于欧洲计算机视觉会议，第 361–379 页。Springer，2024 年。3[73] 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》。arXiv 预印本 arXiv:2506.09982，2025 年。2、6、7、13、14、15、18[74] Fanbo Xiang、Yuzhe Qin、Kaichun Mo、Yikuan Xia、Hao Zhu、Fangchen Liu、Minghua Liu、Hanxiao Jiang、Yifu Yuan、He Wang 等。《Sapien：一个仿真部件交互环境》。发表于 IEEE/CVF 计算机视觉与模式识别会议论文集，第 11097–11107 页，2020 年。2[75] Yiming Xie、Chun-Han Yao、Vikram Voleti、Huaizu Jiang 和 Varun Jampani。《Sv4d：具有多帧和多视图一致性的动态 3D 内容生成》。arXiv 预印本 arXiv:2407.17470，2024 年。2[76] Jinbo Xing、Menghan Xia、Yong Zhang、Haoxin Chen、Wangbo Yu、Hanyuan Liu、Gongye Liu、Xintao Wang、Ying Shan 和 Tien-Tsin Wong。《DynamiCrafter：利用视频扩散先验为开放域图像添加动画》。发表于欧洲计算机视觉会议，第 399–417 页。Springer，2024 年。6、18[77] Zhen Xu、Sida Peng、Chen Geng、Linzhan Mou、Zihan Yan、Jiaming Sun、Hujun Bao 和 Xiaowei Zhou。《从稀疏视图视频生成可重照明、可动画的神经化身》。发表于 IEEE/CVF 计算机视觉与模式识别会议论文集，第 990–1000, 2024 – 页。Zeyu [78] Yang、Zijie Pan、Chun Gu 和 Li Zhang。《Diffusion^2：通过视频和多视图扩散模型的分数合成生成动态 3D 内容》。arXiv 预印本 arXiv:2404.02148，2024 年。2[79] Chun-Han Yao、Yiming Xie、Vikram Voleti、Huaizu Jiang 和 Varun Jampani。《Sv4d 2.0：增强多视图视频扩散中的时空一致性，以实现高质量 4D 生成》。arXiv 预印本 arXiv:2503.16396，2025 年。2[80] Vickie Ye、Ruilong Li、Justin Kerr、Matias Turkulainen、Brent Yi、Zhuoyang Pan、Otto Seiskari、Jianbo Ye、Jeffrey Hu、Matthew Tancik 和 Angjoo Kanazawa。《gsplat：用于高斯泼溅的开源库》。机器学习研究期刊，26(34)：1–17，2025 年。13[81] Yufei Ye、Abhinav Gupta、Kris Kitani 和 Shubham Tul- siani。《G-hop：用于交互重建和抓取合成的生成式手物先验》。发表于 IEEE/CVF 计算机视觉与模式识别会议论文集，第 1911–1920 页，2024 年。3[82] Jiraphon Yenphraphai、Ashkan Mirzaei、Jianqi Chen、Ji- axu Zou、Sergey Tulyakov、Raymond A Yeh、Peter Wonka 和 Chaoyang Wang。《Shapegen4d：迈向从视频生成高质量 4D 形状》。arXiv 预印本 arXiv:2510.06208，2025 年。2[83] Heng Yu、Chaoyang Wang、Peiye Zhuang、Willi Mena- pace、Aliaksandr Siarohin、Junli Cao、L´aszl´o Jeni、Sergey Tulyakov 和 Hsin-Ying Lee。《4real：通过视频扩散模型迈向照片级真实感的 4D 场景生成》。神经信息处理系统进展发表于神经信息处理系统，37 45280，2024 45256 2：–。[84] Mark Yu、Wenbo Hu、Jinbo Xing 和 Ying Shan。Trajectorycrafter：通过扩散模型为单目视频重新定向相机轨迹。载于 IEEE/CVF 国际计算机视觉会议（ICCV）论文集，第 100–111 页，2025。6、7、13、14、15、18[85] Yu-Jie Yuan、Leif Kobbelt、Jiwen Liu、Yuan Zhang、Pengfei Wan、Yu-Kun Lai 和 Lin Gao。4dynamic：结合混合先验的文本到4D生成。arXiv 预印本 arXiv:2407 12684，2024 2。。[86] 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 IEEE/CVF 国际计算机视觉会议论文集，第 12502–12513 页，2025。2[87] Tianyuan Zhang、Hong-Xing Yu、Rundi Wu、Brandon Y Feng、Changxi Zheng、Noah Snavely、Jiajun Wu 和 William T Freeman。Physdreamer：通过视频生成与3D物体进行基于物理的交互。载于欧洲计算机视觉会议论文集，第 388–406 页。Springer，2024。3[88] Silvia Zuffi、Angjoo Kanazawa、David Jacobs 和 Michael J. Black。3D动物园：动物3D形状与姿态建模。载于 IEEE 计算机视觉与模式识别会议（CVPR），2017。2

A. 实现细节

A.1. 管线实现细节

我们实验中使用的3D 资产从 Sketchfab 下载 [1] 和 BlenderKit [7], , 并在 Blender 中构建静态场景快照 [11]。用于 [31] 基于网格的初始化和 4D 优化的 3D-GS 渲染均使用`gsplat` [80]。我们采用 Wan 2.2（14B）图像转视频模型 [65]作为视频生成模型。所有训练均以 832×464 的分辨率进行（Wan 2.2 的默认设置），并对41 帧的变形序列进行优化。

控制点根据占据网格的中心点进行初始化。具体而言，对于每个物体，我们首先根据其给定的网格计算有符号距离场（SDF） $\phi_i(\mathbf{x})$。然后，我们提取物体内部体素中心的集合，记为 $\mathcal{I}_i = \{\mathbf{x} \mid \phi_i(\mathbf{x}) \leq 0, \mathbf{x} \in V_s\}$，其中 V_s 表示体素尺寸为 s的网格中的所有体素中心点。最后，我们对 $\mathcal{I}_i$ 先进行最远点采样，再进行 K 均值聚类，以确定控制点的位置 $\mathbf{p}_k$。此外，我们将每个控制点协方差矩阵 $\mathbf{\Sigma}_k$ 中的尺度初始化为其与三个最近邻控制点之间距离的平均值，并将初始旋转设为单位变换。为确保优化稳定，我们固定$\mathbf{p}_k$，仅在训练过程中优化 $\mathbf{\Sigma}_k$。在形变训练中，我们还引入了分阶段训练计划：在第 100 次迭代时，将所有第 30 次迭代之后的形变重新初始化为第 30 次迭代时的形变，从而进一步促进后续帧的稳定学习。

我们采用 3D-GS 中使用的对数线性学习率计划。存储在芬威克树中的变形的学习率从 0.006 衰减至 0.00006。控制点尺度的学习率采用相同的衰减方式（从 0.006 衰减至 0.00006），而旋转的学习率则从 0.003 衰减至 0.00003。CFG [23] 引导强度从 25 线性衰减至 12。时间正则化损失的权重从 9.6 衰减至 1.6，空间正则化损失的权重则从 3000 衰减至 300。用于提取时间正则化中均匀分布的点云以及初始化控制点的体素尺寸 s 通过二分搜索自动确定，使得表面附近的体素中心数量满足 $|\mathcal{S}_i| \approx 7500$。每个资产以批次大小为 4 训练 2,000 次迭代，在 NVIDIA H200 GPU 上耗时约 20 小时。

A.2. 机器人操作实现细节

对于用于生成稠密物体流的物体，在“拿起香蕉”和“放下台灯”案例中，我们直接扫描了真实物体，并将扫描结果输入我们的流程。对于其他案例，由于难以准确扫描物体，我们改为测量其长度统计数据，并在 Blender 中创建尺寸相符的数字孪生体，然后将其输入我们的流程。

A.3. 基线实现细节

对于 Animate3D [30] 和 AnimateAnyMesh [73], , 我们将场景中的所有物体合并为一个网格，并直接输入其流程。对于 MotionDreamer [64], , 我们遵循其设置，使用神经雅可比场（NJF） [2] 作为动画模型，并为每个物体分别训练一个 NJF。为了对从 TrajectoryCrafter [84],采样的视频进行稳健的 4D 重建，我们使用一组粗略的控制点，并采用基于 Fenwick 树的形变序列作为 4D 表示。此外，我们还在优化过程中同时应用时间正则化损失和空间正则化损失。

B. Rectified Flow 模型的 SDS 推导

当从 $\mathcal{U}(0,\ 1)$ 中均匀采样噪声级别 τ 时，Rectified Flow（RF）模型的训练损失为：

$$\mathcal{L}_{\text{RF}}(\theta; \mathbf{z}, \mathbf{y}) = \mathbb{E}_{\tau\sim\mathcal{U}(0,1),\,\epsilon}\left[w(\tau)\left\|\hat{v}(\mathbf{z}_\tau; \tau, \mathbf{y}) - (\epsilon - \mathbf{z})\right\|^2\right], \tag{14}$$

其中，$\epsilon \sim \mathcal{N}(0,\ I)$，而 $\mathbf{z}_\tau = (1-\tau)\mathbf{z} + \tau\epsilon$ 是线性插值后的潜变量。

对 $\mathcal{L}_{\text{RF}}$ 求 $\mathbf{z}$ 的导数可得：

$$\nabla_{\mathbf{z}}\mathcal{L}_{\text{RF}}(\theta; \mathbf{z}, \mathbf{y}) = \mathbb{E}_{\tau\sim\mathcal{U}(0,1),\,\epsilon}\left[w(\tau)\left(\hat{v}(\mathbf{z}_\tau; \tau, \mathbf{y}) - (\epsilon - \mathbf{z})\right)\left(\frac{\partial\hat{v}(\mathbf{z}_\tau; \tau, \mathbf{y})}{\partial\mathbf{z}} + I\right)\right]. \tag{15}$$

遵循分数蒸馏采样（SDS） [53], 的推导风格，我们省略了通过 RF 模型反向传播的项，$^{\partial}{}_{\hat{v}(\mathbf{z}_\tau;\tau,\mathbf{y})\,\partial\mathbf{z}}$，并应用链式法则，从~~$\mathbf{z}$ 反向~~传播至 4D 表示的参数 θ。由此得到正文中使用的 RF-SDS 梯度：

$$\nabla_\theta\mathcal{L}_{\text{RFSDS}}(\theta; \mathbf{z}, \mathbf{y}) = \mathbb{E}_{\tau,\,\epsilon}\left[w(\tau)\left(\hat{v}(\mathbf{z}_\tau; \tau, \mathbf{y}) - (\epsilon - \mathbf{z})\right)\frac{\partial\mathbf{z}}{\partial\theta}\right]. \tag{16}$$

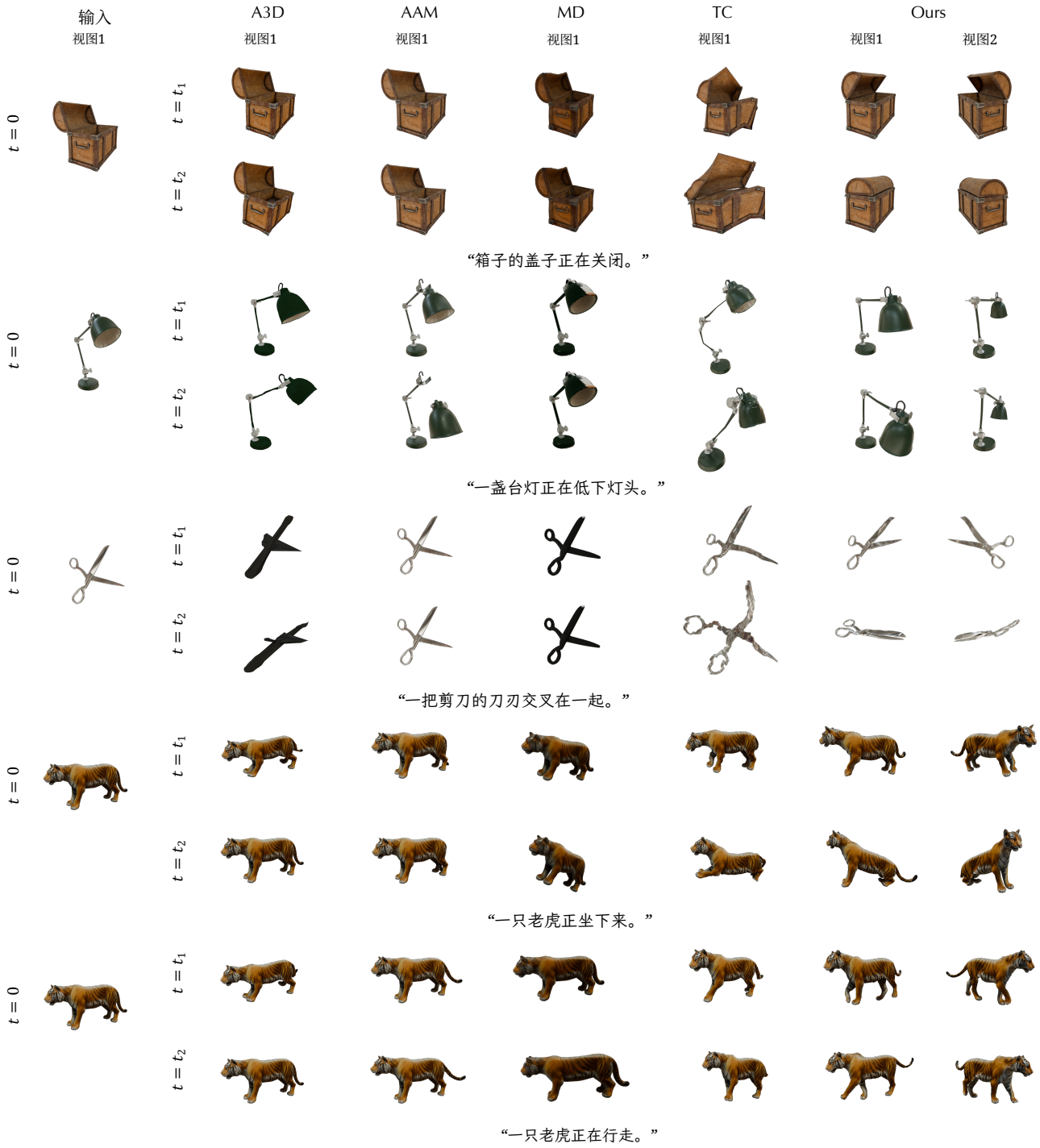

图 11. **单网格动画的定性比较**。 我们将我们的方法与几种网格动画方法进行了比较。我们的方法生成的结果与给定提示更加一致，运动也更加自然。图中，A3D 指 Animate3D [30], AAM 指 AnimateAnyMesh [73], MD 指 MotionDreamer [64], TC 则表示由 TrajectoryCrafter 根据视频生成的 4D重建结果 [84]。

C. 更多实验结果

本节中，我们展示了该方法的更多实验结果。

指标	物体	Animate3D	AnimateAnyMesh	MotionDreamer（Orig）	MotionDreamer（Wan）	TC	Ours	总计
提示词对齐	Cat	0	2	0	1	0	96	99
	Dog	0	3	0	1	7	88	99
	拥抱	1	0	0	0	1	97	99
	机器人	0	0	1	1	2	95	99
	海狮	1	0	1	2	43	52	99
	砖块	0	1	1	0	4	93	99
	Avg	0.3333	1	0.5	0.8333	9.5	86.8333	99
运动真实感	Cat	0	0	1	1	2	95	99
	Dog	1	2	1	0	11	84	99
	拥抱	0	0	0	0	3	96	99
	机器人	0	1	1	1	1	95	99
	海狮	1	0	2	0	37	59	99
	砖块	1	0	0	0	8	90	99
	Avg	0.5	0.5	0.8333	0.3333	10.3333	86.5	99

表 2. 关于生成场景级 4D 运动的用户研究原始结果。我们展示了每位参与者针对特定指标认为最佳选项的投票数。

指标	物体	Animate3D	AnimateAnyMesh	MotionDreamer（Orig）	MotionDreamer（Wan）	TC	Ours	总计
提示词对齐	箱子	2	0	1	0	0	47	50
	Lamp	0	1	2	1	0	46	50
	剪刀	1	0	1	1	0	47	50
	坐着	4	1	1	0	5	39	50
	行走	1	0	3	0	1	45	50
	平均值（原始）	1.6	0.4	1.6	0.4	1.2	44.8	50
运动真实感	箱子	2	0	1	1	1	45	50
	Lamp	5	0	2	0	0	43	50
	剪刀	0	1	7	0	1	41	50
	坐着	5	1	2	0	6	36	50
	行走	2	2	1	0	0	45	50
	平均值（原始）	2.8	0.8	2.6	0.2	1.6	42	50

表 3. 单物体4D运动生成定量比较的用户研究结果。

C.1. 单网格动画比较

我们进一步将我们的方法与基线方法在单物体网格动画任务上进行比较。基线方法与论文正文中相同：Animate3D [30], AnimateAnyMesh [73], MotionDreamer [64], 以及基于 TrajectoryCrafter 生成的视频进行的4D重建 [84]。我们在五个提示词上评估所有方法："箱子的盖子正在关闭""台灯正在低下灯头""一把剪刀的刀刃交叉在一起""一只老虎坐下""一只老虎正在行走"。

定性结果如图 11所示。与现有方法相比，我们的方法始终能实现更好的提示词对齐，并生成更自然的运动。在定量评估中，我们邀请了 50 位参与者开展用户研究，将我们的结果与所有基线方法进行比较：**89.6%** 的参与者认为我们的方法在提示词对齐方面表现最佳，另有 **84%**的参与者认为我们的方法在运动真实感方面表现最佳。这些结果进一步体现了我们的方法相较于现有方法的优势。完整结果见表3。

C.2. 用户研究完整表格

在表 2 和表 3中，我们提供了完整的用户研究结果，包括每个场景中偏好各方法的参与者人数。在所有场景中，我们的方法在提示词对齐和运动真实感方面都获得了最高偏好。

C.3. 失败案例

我们的失败案例主要源于两个因素：（1）底层视频生成模型的局限性；（2）无法处理在静态快照中不存在、但随后出现在运动序列中的物体。示例见图 12。我们的失败案例主要源于两个因素：（1）底层视频生成模型的局限性；（2）无法处理在静态快照中不存在、但随后出现在运动序列中的物体。示例

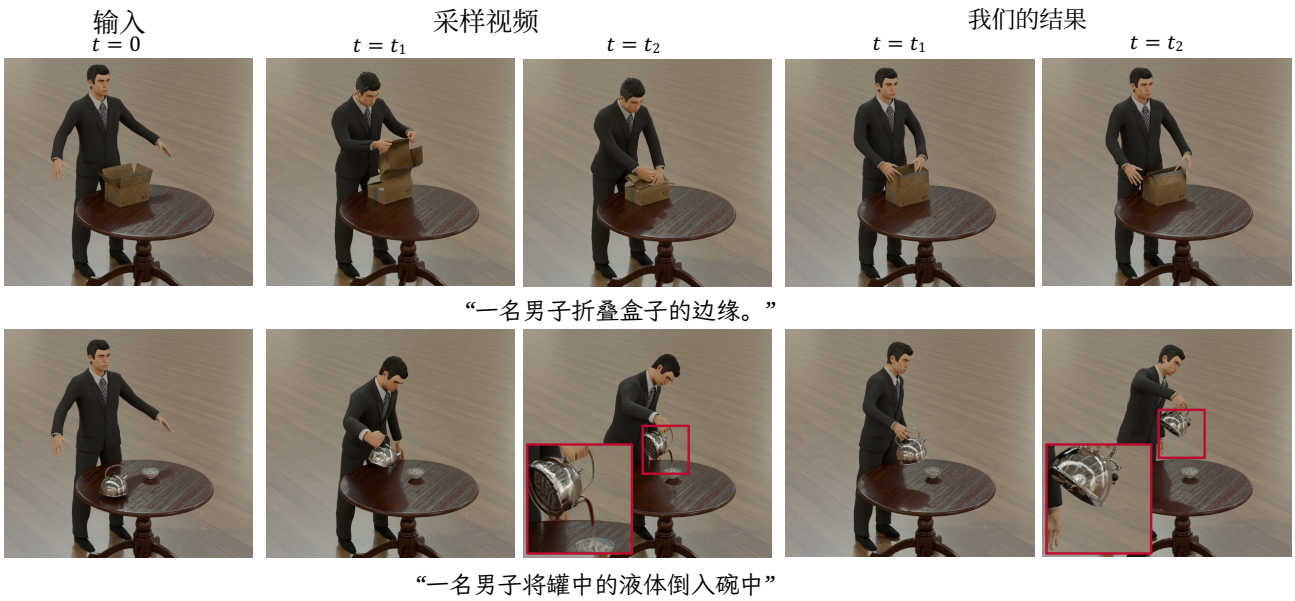

图 12。**失败案例**。 第一行的失败源于视频生成模型的局限：它无法生成与提示相符的运动，这一点从它无法采样出与所描述动作一致的视频便可看出。第二行的失败则是因为我们的方法无法生成初始静态场景中不存在的物体。因此，即使提示中要求出现液体，液体也不会出现，因为系统无法生成新出现的物体。

如图 12所示。下面将详细说明。

视频生成模型的局限性。 由于我们的方法是从预训练视频生成模型中蒸馏而来，其能力本质上取决于底层模型。如果生成器无法合成与提示一致的视频，我们的 4D 优化就会收到误导性的梯度。在这种情况下，我们的方法无法生成正确的运动。如图 12第一行所示，视频模型反复无法采样出与提示一致的视频，导致我们的方法生成错误的运动。

无法处理新出现的物体。 我们方法的另一项局限在于，无法处理初始静态快照中不存在的物体。我们的 4D 表示只能使开始时已有的几何发生形变，因此无法创建序列后续阶段本应出现的物体。当提示涉及新物体进入场景时，监督信号要求生成系统无法实现的运动。在这些情况下，优化过程要么忽略所请求的效果，要么产生不完整的运动，如图 12第二行所示；由于系统无法引入新的几何，因此没有液体出现。

D. 局限性与未来工作

尽管我们的方法能够生成包含多个物体之间高度逼真交互的动态场景，但仍存在若干局限，也为未来工作指明了有前景的方向。对于第 C.3节所述的失败案例，由底层视频生成模型局限性导致的问题，可能会随着视频生成技术的持续进步而得到缓解。对于因初始静态场景中不存在新出现的物体而导致的失败，一种潜在的解决方案是在优化过程中加入能够生成新几何的模块。

除了失败案例之外，我们方法的另一项局限是训练时间较长。根据我们的观察，大部分运行时间都耗费在通过变分自编码器进行反向传播上 [33]。一个有前景的未来方向是开发一种蒸馏策略，从而完全避免通过变分自编码器进行反向传播。这或许可行，因为我们的目标是生成运动，而非RGB外观，这表明为实现有效的运动监督，可能并不严格需要完整的变分自编码器梯度。

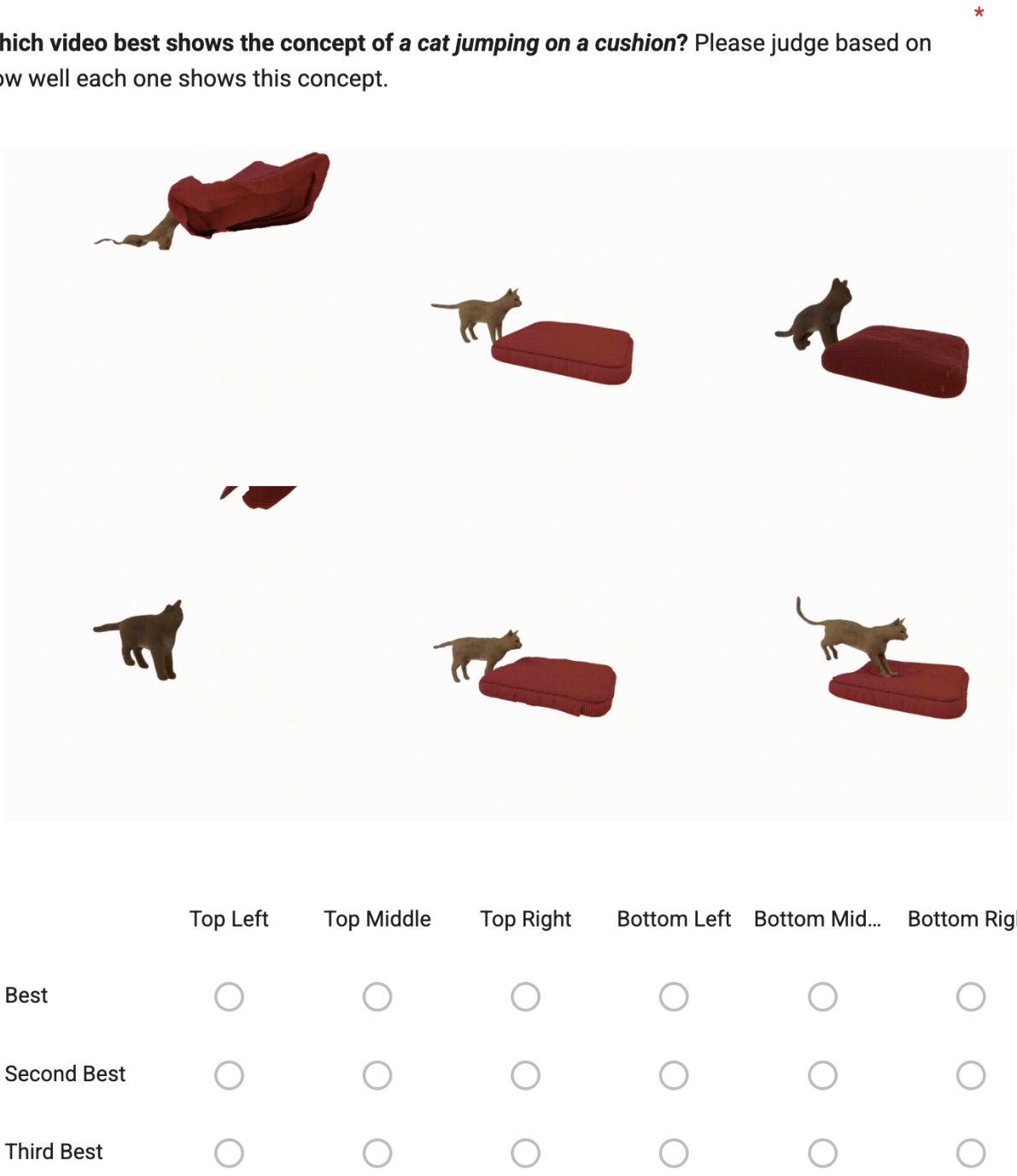

图13。 **关于提示词对齐的用户研究问题截图。**

E. 用户研究模板

我们在图13 和图14中提供了用户研究界面的截图。我们要求参与者从五种方法中选出最佳、次佳和第三佳结果。从左到右、从上到下，对应的方法

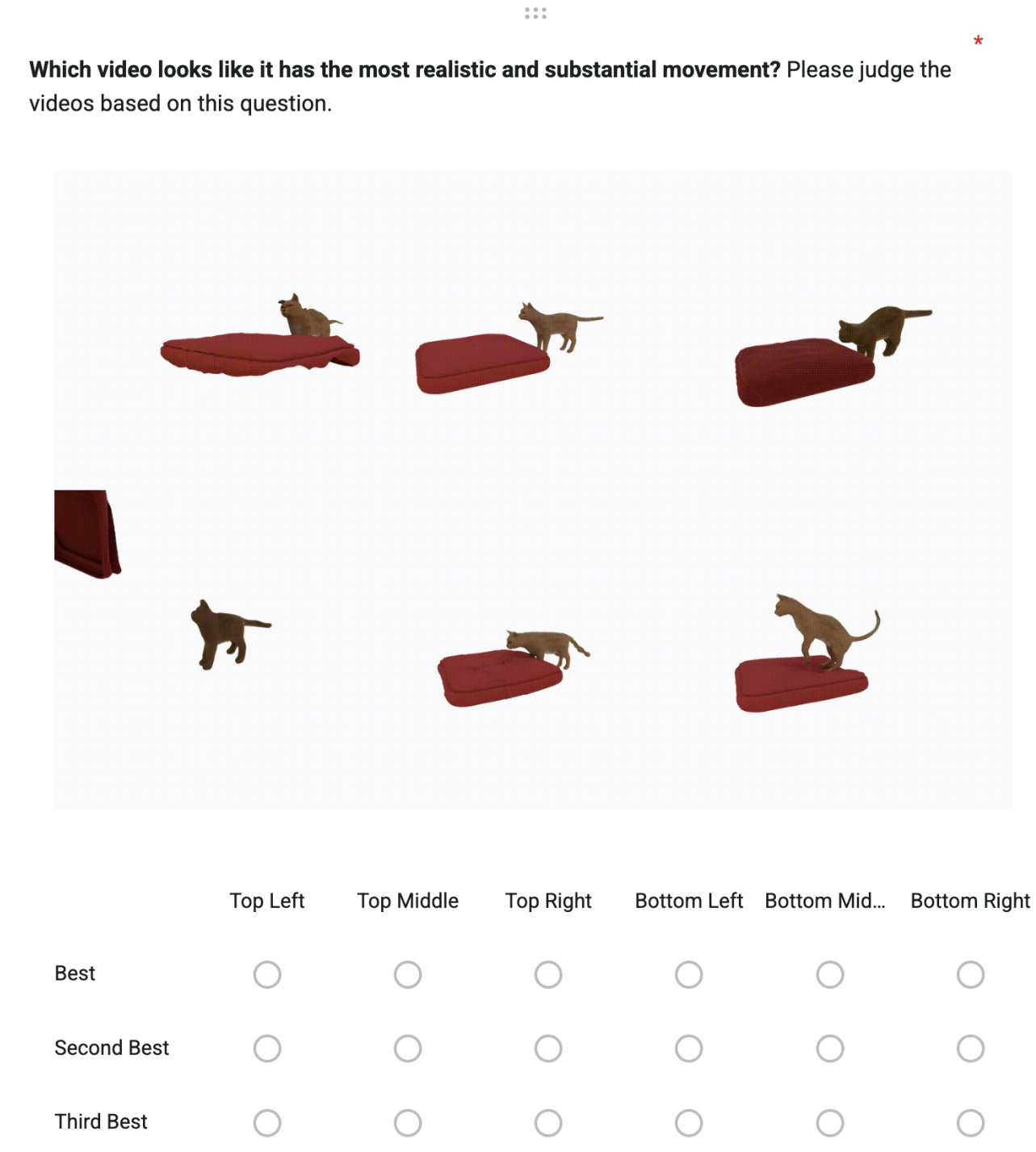

图14. **关于运动真实感的用户研究问题截图。**

包括：Animate3D [30], AnimateAnyMesh [73], MotionDreamer [64] 使用 DynamiCrafter [76], 使用 Wan 2 的 MotionDreamer。2 [65], 4由 TrajectoryCrafter 生成的视频进行 D 重建 [84], 以及我们的方法。