DIPO: Dual-State Images Controlled Articulated Object Generation Powered by Diverse Data

Ruiqi Wu[1, 2, 3*] **Xinjie Wang**[3] **Liu Liu**[3] **Chunle Guo**[1, 2] **Jiaxiong Qiu**[3]
Chongyi Li[1, 2†] **Lichao Huang**[3] **Zhizhong Su**[3] **Ming-Ming Cheng**[1, 2]
[1]NKIARI, Shenzhen Futian [2]VCIP, CS, Nankai University [3]Horizon Robotics

Abstract

我们提出了 **DIPO**，这是一个用于从图像对（一张处于静止状态，另一张处于活动状态）可控地生成铰接式 3D 物体的新颖框架。与单图像方法相比，我们的双图像输入仅增加了少量的数据收集开销，但同时提供了重要的运动信息，为预测部件间的运动学关系提供了可靠指导。具体而言，我们提出了一个双图像扩散模型，它能捕捉图像对之间的关系，以生成部件布局和关节参数。此外，我们引入了一个基于思维链 (CoT) 的**图推理器** (graph reasoner)，用于显式推断部件的连接关系。为了进一步提高模型在复杂铰接物体上的鲁棒性和泛化能力，我们开发了一个名为 **LEGO-Art** 的全自动数据集扩展流水线，用以丰富 PartNet-Mobility 数据集的多样性和复杂性。我们提出了 **PM-X**，这是一个大规模的复杂铰接 3D 物体数据集，并附带了渲染图像、URDF 标注和文本描述。大量实验表明，无论是在静止状态还是活动状态下，DIPO 的性能都显著优于现有的基线方法；同时，我们提出的 PM-X 数据集进一步增强了模型对多样化且结构复杂的铰接物体的泛化能力。我们的代码和数据集已发布于 https://github.com/RQ-Wu/DIPO。

1 引言

铰接物体在日常环境中无处不在。实现铰接结构的精确建模是构建交互式虚拟环境的关键。它在模拟 [49, 42, 46]、动画 [48, 5, 33, 21]、机器人操控 [11, 9, 28, 32] 和具身智能 (embodied AI) [19, 36, 31, 20, 16] 等领域扮演着至关重要的角色。

*This work was done while Ruiqi Wu was a Research Intern with Horizon Robotics.
†denotes correspondence author.

Preprint. Under review.

然而，手动构建此类模型不仅劳动强度大，而且可扩展性差。因此，越来越多的研究致力于开发自动化的铰接物体建模方法 [39, 22, 43, 18, 24, 6, 23]。尽管现有方法已经取得了积极的进展，但在应用于结构复杂或视觉模糊的物体时，它们的性能会明显下降。这些局限性源于两个根本瓶颈。

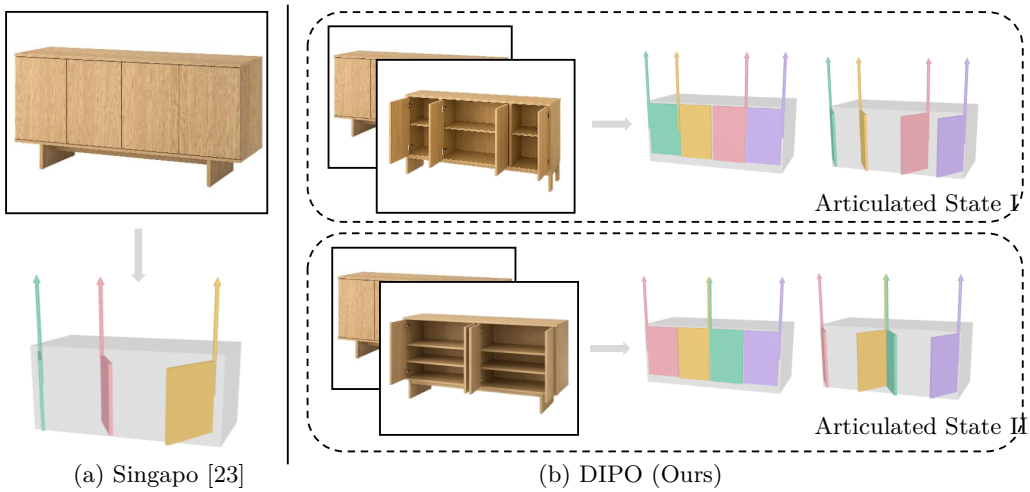
第一个问题是 **输入模态的约束** (input modality constraints)。基于重建的方法 [39, 22, 43] 通常依赖多视图或多状态的图像来高精度地重建铰接行为。这些方法虽然有效，但需要昂贵的数据采集设备、精确的相机标定和严格对齐的时序输入，因此难以规模化。受益于扩散模型的可控性 [12, 37, 35, 30, 45, 51, 50]，许多基于生成的方法 [18, 24, 6, 23] 被提了出来作为另一条研究路线。它们利用最小化的输入（例如类别先验或单张 RGB 图像）来直接合成铰接物体。然而，类别先验缺乏空间特异性，而单图像输入则缺乏明确的铰接信息。因此，这些方法只能以概率的方式推断运动学行为。最终，这两类方法在面对具有挑战性的数据时，都无法同时提供可控性和泛化能力。

其次是 **训练数据的局限性** (limitations in training data)。数据驱动的建模方法需要兼具铰接多样性和结构复杂性的大规模数据集。然而，大多数现有数据集都在某些方面存在不足。例如，PartNet-Mobility (PM) [46] 提供了大量的铰接实例，但其实例以简单和重复的布局为主，可变性有限。相比之下，Articulated Container Dataset (ACD) [14] 包含更真实、结构更多样的物体，但其规模较小，限制了其在模型训练中的应用价值。

为了解决第一个问题，我们提出了 **DIPO**，这是一个基于静止（闭合）状态和活动（打开）状态图像对作为条件来生成 3D 铰接物体的框架。双状态图像对编码了关键的运动线索和连接信息。与单图像方法相比，双状态输入解决了部件运动和空间关系中的模糊性。而相较于多视图方法，它在保持足够铰接信息的同时，采集难度显著降低。DIPO 构建于一个扩散型 Transformer 架构 [30] 之上，包含两个核心组件。首先，一个 *双状态注入模块* (Dual-State Injection Module) 帮助网络建模双状态图像之间的关系。其次，一个基于思维链 (CoT) 技术 [41, 17] 的 *图推理器* (Graph Reasoner) 按部就班地推断部件的连接关系。此外，该模块在由 GPT-4o [3, 1] 合成的视觉提示上进行少样本学习 (few-shot learning)，以获得更好的性能。所提出的方法在铰接 3D 物体的生成中实现了更高的可控性和更好的性能。

针对第二个挑战，我们提出了一个名为 **PartNet-Mobility-Complex (PM-X)** 的新数据集，它提供了多样化且结构复杂的铰接物体，并附带了渲染图像、URDF 标注 [34] 和语言描述。PM-X 是通过一个基于智能体系统 (agent system) 的全自动数据构建流水线（名为 **LEGO-Art**）构建的。该流水线从一个大型语言模型 (LLM) [3] 采样的自然语言提示开始，首先在离散化的 3D 空间中生成粗略的部件布局。然后，我们开发了一个工具包，将它们转换为具有精确坐标和铰接参数的标注。基于检索算法 [24]，我们可以获得最终的 3D 物体和渲染图像。最后，使用一个视觉语言模型 (VLM) [1] 来过滤掉不合理的样本。

我们从互联网收集一张静止状态的图像，并使用视觉生成模型 [1] 生成相应的活动状态图像。如图 1 所示，我们的方法优于当前最先进的方法，即 SINGAPO [23]。



(a) Singapo [23] (b) DIPO (Ours)

图 1: **真实采集数据的视觉对比**。(a) SINGAPO 难以处理具有挑战性的数据，并且由于依赖单一输入而无法建模运动关系。然而，我们的 DIPO (b) 以双状态图像对为条件，能有效生成准确的布局，并实现对不同活动状态下部件运动的精确控制。

我们的主要贡献总结如下：

- 我们提出了一种新颖的双状态图像模型，用于可控地生成铰接 3D 物体，该模型融合了布局扩散和基于 CoT 的连接关系推理。
- 我们开发了 LEGO-Art 流水线来构建结构多样的铰接物体，并贡献了一个新的大规模数据集 PM-X，该数据集包含渲染图像和物理标注。
- 大量实验表明，DIPO 的性能显著优于最先进的方法，并且我们提出的 LEGO-Art 和构建的 PM-X 数据集增强了模型对复杂结构的泛化能力。

2 相关工作

2.1 铰接式物体创建

铰接物体建模的最新进展大致可分为基于重建 (reconstruction-based) 和基于生成 (generation-based) 的两大类方法。

重建方法 (Reconstruction methods) 通常依赖多视图或多状态输入来重建部件级的几何形状和铰接参数。CLA-NeRF [39] 在已知类别内，从稀疏的多视图 RGB 图像中重建铰接物体。PARIS [22] 利用双状态的多视图 RGB 图像，将这一设定扩展到了未知类别。Weng 等人 [43] 进一步引入了深度信息，以支持更丰富的几何先验。然而，这些方法依赖于密集对齐的输入和已知的部件数量，限制了它们在现实世界场景中的适用性。相比之下，我们的方法仅以一对图像为条件，这降低了输入复杂性，同时保持了铰接的保真度。

生成方法 (Generative approaches) 旨在从紧凑的输入中合成铰接物体，从而绕过了对密集观测的需求。NAP [18] 将布局和铰接参数解析为图，并无条件地生成铰接 3D 物

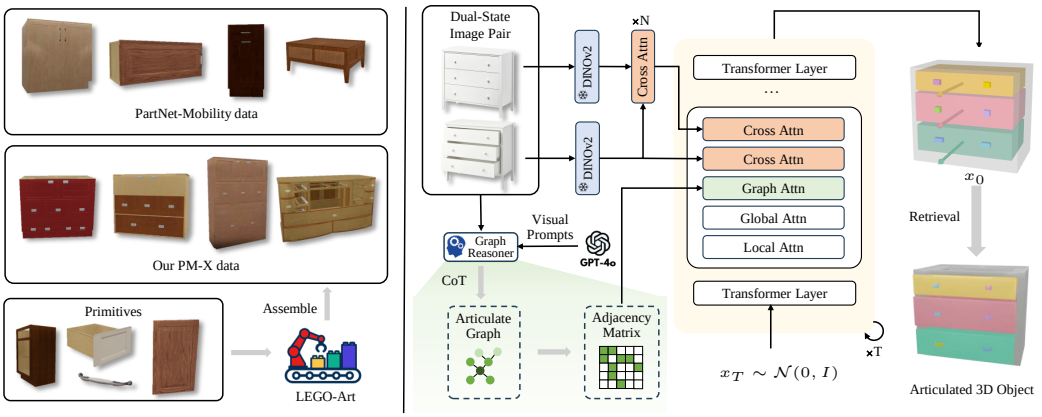


图 2: 所提出的 DIPO 框架概览。左侧部分展示了所提出的 LEGO-Art 流水线，它组装现有数据集中的基元 (primitives) 来构建 PM-X 数据集，该数据集比 PM 数据集更多样、更复杂。右侧部分展示了我们的扩散模型，它配备了基于 CoT 的图推理器 (Graph Reasoner) 来进行铰接图推理，并以静止和活动图像对为条件来生成铰接物体。

体。CAGE [24] 实现了基于给定铰接图的可控生成。尽管这些模型支持高效采样，但它们缺乏明确的视觉指导来实现更准确的可控性。URDFormer [6] 通过结合一个视觉检测器 [25, 44] 来提取空间布局和一个 Transformer 来预测铰接参数，解决了这个问题。SINGAPO [23] 提出了一个以静止状态图像为条件的扩散模型 [12, 37, 35, 30] 来生成铰接物体。然而，由于缺乏明确的铰接动力学信息，现有方法的可控性仍然有限。所提出的 DIPO 通过利用在静止和活动状态下捕获的一对图像所提供的运动信息，有效地解决了这一局限性。

2.2 合成铰接对象数据集

具备部件级结构的大规模 3D 数据集的出现，极大地推动了铰接物体建模的研究。早期的数据集，例如 [13, 47] 中使用的数据集，是通过从 ShapeNet [4] 和 SketchUp [38] 中手动分割形状，并为部件对标注铰接参数来构建的。Shape2Motion [40] 引入了一个支持通过动画进行视觉验证的标注工具，从而扩大了数据集的规模。PartNet-Mobility[46] 是一个基于 PartNet[27] 构建的大规模铰接物体数据集。它提供了部件级铰接的标注以及高质量的渲染图像，是目前应用最广泛的基准之一。GAPartNet [10] 专注于跨类别的功能性部件检测，强调可泛化和可操作的部件，如按钮和把手。这些数据集推动了用于铰接分析的深度学习模型的发展，但在结构复杂性和多样性方面仍然有限。为了提高铰接的多样性和真实感，ACD [14] 从 ABO [7]、3D-Future [8] 和 HSSD [15] 收集了复杂的铰接物体。虽然 ACD 中的铰接结构更为复杂，但数据集的规模仍然有限。为了同时解决多样性和可扩展性的限制，我们提出了 PM-X，这是一个大规模、兼容 URDF 且具有高结构复杂性的程序化生成铰接物体数据集。

3 从双图像对生成铰接物体

3.1 概述

我们提出了一个扩散网络，用于以一对双状态图像和部件级连接图为条件，生成铰接物体的所有参数。整体架构如图 2 所示。为了支持这一生成过程，我们对每个部件的空间位置、铰接连接性以及语义属性进行了参数化。第 i 个部件 $\mathbf{p}_i$ 由边界框坐标 $\mathbf{b}_i \in \mathbb{R}^6$、语义标签 l_i、铰接类型 t_i、关节轴 $\mathbf{a}_i \in \mathbb{R}^6$ 和运动范围 $\mathbf{r}_i \in \mathbb{R}^2$ 来表示。为了方便统一处理，所有属性都被重复为一个 6 维数组，最终形成每个部件的 5×6 矩阵表示。

3.2 双状态图像条件化

我们对去噪过程进行静止状态和活动状态图像的条件化，以捕获运动感知线索。令 $\mathcal{F}_R$ 和 $\mathcal{F}_A$ 分别表示来自静止图像和活动图像的 DINOv2 [29] 特征。为了将这些特征集成到扩散网络中，我们在每一层应用一个 **双状态注入模块** (Dual-State Injection Module, DIM)。

(a) 静止状态　　(b) 活动状态

图 3: 图推理器 (Graph Reasoner) 使用的双状态视觉提示。GPT-4o 可以生成逼真且结构复杂的图像对。

给定部件嵌入 X，我们首先与静止状态特征 $\mathcal{F}_R$ 进行交叉注意力 (cross-attention) 操作以捕获静态外观。然后，我们引导活动状态特征 $\mathcal{F}_A$ 关注 $\mathcal{F}_R$，随后将这一经过上下文增强的信号注入到 X 中。在每个扩散步骤中，整体的条件化更新定义为：

$$X = X + \mathrm{CA}(X, \mathcal{F}_R) + \mathrm{CA}(X, \mathrm{CA}(\mathcal{F}_A, \mathcal{F}_R)), \tag{1}$$

其中 $\mathrm{CA}(Q, K)$ 表示一个标准的交叉注意力操作，即查询 Q 关注键值源 K。这种设计允许模型通过对比两种输入状态，生成更准确的部件运动和关节行为。

3.3 基于思维链提示的图推理器

我们引入了 **图推理器** (Graph Reasoner)，这是一个基于思维链 (Chain-of-Thought, CoT) 的模块，它根据双状态图像预测铰接部件的连接图，作为扩散过程的结构先验。推理过程遵循一个分步进行的范式。它首先识别候选部件并估计其粗略的空间布局，然后验证该布局是否满足给定的铰接规则，最后推断出连接关系以生成铰接图。接着，我们将预测的铰接图转换为一个邻接矩阵 (adjacency matrix)，该矩阵作为注意力掩码 (attention mask)，引导扩散模型的自注意力沿着有效的结构连接进行。

此外，我们利用 GPT-4o 的指令遵循和视觉编辑能力，生成结构多样的物体的双状态图像对，如图 3 所示。这些结果作为图推理器的有力示例视觉提示 (visual prompts)，有助于提高图预测的稳定性和泛化能力。

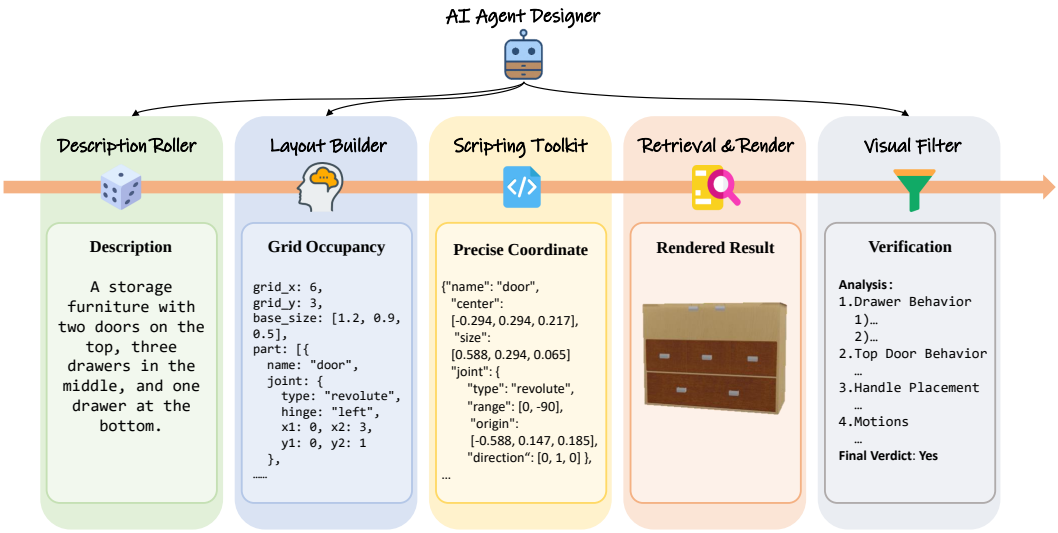


图 4: 所提出的 PM-X 数据集的全自动合成流水线概览。该合成流水线由五个按顺序执行的功能模块组成：(1) 描述生成器 (description roller)，使用一个大型语言模型 (LLM) 为结构化布局生成自然语言描述，(2) 布局构建器 (layout builder)，用于生成部件级的网格占用和关节配置，(3) 脚本工具包 (scripting toolkit)，用于从基于网格的布局信息构建精确坐标，(4) 检索与渲染模块 (retrieval and render module)，用于组装几何体并渲染双状态图像，以及 (5) 视觉过滤器 (visual filter)，使用一个视觉语言模型 (VLM) 来验证生成样本的合理性。特别地，模块 (1)、(2) 和 (5) 由 AI 智能体设计器 (AI Agent Designer) 自动构建和管理。

4 从 PartNet-Mobility 构建复杂数据

4.1 LEGO-Art 流水线

为了在具有挑战性的数据上获得良好性能，我们需要一个具有多样化部件布局的大规模 3D 数据集。然而，现有数据集在不同方面仍存在不足：PM [46] 提供了充足的数据，但缺乏铰接复杂性；而 ACD [14] 包含了更真实的运动学结构，但数据集规模有限。

为了解决这个问题，我们设计了一个 **L**anguage-driven **E**ngine via **G**rid **O**rganization for **Art**iculation objects construction (**LEGO-Art**，基于语言驱动和网格组织的铰接物体构建引擎)。这是一个全自动的合成流水线，通过组装现有数据集中的部件基元 (part primitives) 来生成复杂的铰接 3D 资产。图 4 展示了该合成流水线的整体工作流程。下面将详细说明每个步骤。

- **描述生成器 (Description Roller)**。该流水线首先由一个大型语言模型 (LLM) 智能体生成铰接物体的自然语言描述（例如，“一个储物柜，顶部有两个门，中间有三个抽屉，底部有一个抽屉”）。这为物体的结构提供了一个高层次的蓝图，而无需精确的几何信息。

- **布局构建器 (Layout Builder)**。给定此文本输入，第二个智能体将描述转换为部件布局和铰接配置。我们没有预测精确的 3D 坐标（这常常会引入幻觉），而是将空间离散化为一个网格，并将部件分配到网格单元中。每个部件都关联了关节元数据，如类型、轴向和运动方向。

- **脚本工具包 (Scripting Toolkit)**。我们开发了一个脚本工具包，它将网格级的空间布局转换为精确的 3D 坐标，并分配关节的轴向、方向、运动范围和类型等铰接参数。

- **检索与渲染 (Retrieval & Render)**。我们通过 [24] 提出的算法从 PartNet-Mobility 中检索网格基元 (mesh primitives)，为每个部件分配几何形状。部件根据布局进行缩放和定位，并按照 URDF 的规定进行连接。然后，我们使用 `BLENDER` 为每个物体渲染一对静止和活动状态的图像。

- **视觉过滤器 (Visual Filter)**。为确保数据质量，我们加入了一个最终的过滤步骤。我们使用一个视觉语言模型 (VLM) 来评估每个渲染出的物体是否与其描述合理匹配，以及其铰接是否正确。只有通过此检查的资产才会被包含在我们的最终数据集 PM-X 中。

- **AI 智能体设计器 (AI Agent Designer)**。为了简化上述组件的开发，我们采用了一种基于提示的智能体设计流程。具体来说，我们用自然语言描述了我们预期的系统行为，并使用一个大型语言模型 (LLM) 来共同设计描述生成器、布局构建器和视觉过滤器智能体的系统提示 (system prompts)。

所提出的 LEGO-Art 能够以最少的人力投入，可扩展地生成物理上有效、语义丰富且结构多样的铰接资产，并在使我们的 DIPO 能够泛化到更具挑战性的数据集方面发挥了至关重要的作用。

4.2 PM-X 数据集

基于 LEGO-Art，我们从 PartNet-Mobility 数据集的部件基元构建了一个大规模数据集，命名为 **PM-X**。PM-X 包含 600 个自动生成的结构复杂的铰接物体。对于每个物体，我们进一步提供了相应的渲染图像、URDF 文件和自然语言描述。由于实验设置的原因，我们在所提出的数据集中只考虑了 `StorageFurniture` 和 `Table` 两类物体。然而，该合成流水线可以扩展到更广泛的铰接物体类别，并且整体数据集规模也可以扩大。与现有数据集相比，PM-X 不仅提供了显著更高的结构复杂性和铰接多样性，而且其规模也足以作为一个独立的训练集来训练生成模型。这些特性使其在提高铰接物体生成任务的泛化能力和鲁棒性方面特别有效，尤其是在分布外 (out-of-distribution) 的设置下。我们的实验也证明了 PM-X 数据集的优越性。表 1 表

表 1: 数据集规模与部件复杂度的比较。

数据集	物体数量	平均部件数
PM [46]	570	4.94
ACD [14]	135	7.48
PM-X (Ours)	600	19.40

明，PM-X 数据集在物体数量和平均部件数上都超过了以往的数据集，凸显了其可扩展性和结构丰富性。

5 实验

5.1 实现细节

我们遵循 SINGAPO [23] 的数据集划分方式来构建训练集和测试集。具体来说，训练集由来自 PM [46] 数据集的 493 个铰接物体，以及我们提出的 PM-X 数据集中的 600 个样本组成。每个物体都由 `BLENDER_EEVEE_NEXT` 引擎从 20 个随机视角渲染，以生成双状态图像对。我们进一步引入了复杂的数据增强来提升模型性能，具体细节在补充材料中详述。在评估方面，我们使用了来自 PM 数据集的 77 个留存物体，每个物体从两个随机视角渲染，共得到 144 个双状态测试样本。此外，我们还引入了来自 ACD 数据集 [14] 的 135 个物体，以进一步评估模型对分布外数据的泛化能力。

为了加速收敛，我们使用 CAGE [24] 的预训练权重来初始化我们的模型。我们以 20 的批量大小 (batch size) 训练模型 200 个轮次 (epochs)。该模型使用 AdamW [26] 优化器进行优化，其中 $\beta = (0.9, 0.99)$。图像条件化模块的学习率设置为 5×10^{-4}，基础模型的学习率设置为 5×10^{-5}。所有实验均在 8 块 NVIDIA 4090 GPU 上进行。

5.2 比较

5.2.1 基线方法与评价指标

我们选择了三种有代表性的方法作为比较基线，分别是 URDFormer [6]、NAP [18] 和 SINGAPO [23]。具体来说，为了进行公平比较，我们对预训练的 URDFormer 进行了微调，并重新训练了 SINGAPO。对于 NAP，我们遵循 SINGAPO 的实验设置，在每个层中插入一个图像交叉注意力模块以实现对图像的可控生成，记为 NAP-ICA。

为了评估重建质量和铰接正确性，我们采用了四种评价指标：(1) $d_{\text{gIoU}} \downarrow$，预测部件边界框与真实部件边界框之间的广义交并比 (generalized IoU)；(2) $d_{\text{cDist}} \downarrow$，部件中心之间的欧几里得距离 (Euclidean distance)；(3) $d_{\text{CD}} \downarrow$，预测网格与真实网格之间的倒角距离 (Chamfer Distance) [2]；以及 (4) Acc $\uparrow$，图预测准确率。所有指标都在静止状态和活动状态下进行计算。为清晰起见，我们在表格中用 `RS-` 和 `AS-` 作为指标名称的前缀，以分别表示评估时的状态。

5.2.2 定量比较

我们在表 2 和表 3 中分别报告了在 PM 和 ACD 数据集上的定量结果。为了减少随机波动的影响，我们对每个测试样本使用所有基于扩散的生成方法评估五次，并报告平均指标值。

如表 2 所示，我们的方法 DIPO 在 PartNet-Mobility 测试集上的重建质量和铰接图准确率方面均取得了最佳性能。重要的是，我们观察到，我们的方法从 RS（静止状态）到

表 2: 在 **PartNet-Mobility** 测试集上重建质量和图预测准确率的比较。除了 Acc% (↑) 外，其他指标越低越好 (↓)。

	重建质量						图
	RS-d_{gIoU} ↓	AS-d_{gIoU} ↓	RS-d_{cDist} ↓	AS-d_{cDist} ↓	RS-d_{CD} ↓	AS-d_{CD} ↓	Acc% ↑
URDFormer [6]	1.2327	1.2332	0.2885	0.4403	0.4417	0.6910	6.62
NAP-ICA [18]	0.5706	0.5765	0.0563	0.2547	0.0209	0.3473	25.06
SINGAPO [23]	0.5134	0.5236	0.0487	0.1107	0.0191	0.1270	75.97
DIPO(Ours)	**0.4561**	**0.4683**	**0.0359**	**0.0732**	**0.0132**	**0.0423**	**85.06**

表 3: 在 **ACD** 测试集上重建质量和图预测准确率的比较。除了 Acc% (↑) 外，其他指标越低越好 (↓)。

	重建质量						图
	RS-d_{gIoU} ↓	AS-d_{gIoU} ↓	RS-d_{cDist} ↓	AS-d_{cDist} ↓	RS-d_{CD} ↓	AS-d_{CD} ↓	Acc% ↑
URDFormer [6]	1.1074	1.1094	0.2868	0.3948	0.6229	0.7608	1.52
NAP-ICA [18]	0.9955	1.0000	0.1713	0.3246	0.1141	0.3061	8.27
SINGAPO [23]	0.9700	0.9728	0.1582	0.2057	0.1047	0.1762	36.67
DIPO (Ours)	**0.9126**	**0.9151**	**0.1253**	**0.1541**	**0.0751**	**0.1085**	**48.15**

AS（活动状态）的性能下降幅度明显小于所有其他方法。这表明双图像条件化提供了有效的控制信号，帮助模型保持准确的铰接预测。

在包含更多样化和更真实铰接物体的 ACD 测试集上（表 3），我们的方法继续优于所有基线方法。DIPO 在两种状态下都表现出持续优越的重建精度，并取得了最佳的图预测准确率。在 ACD 数据集上的评估结果表明，我们的方法在分布外数据上表现良好。

以上结果表明，所提出的 DIPO 在结构多样的不同数据集上均实现了优越的定量性能，兼具高准确性和强泛化能力。

5.2.3 定性比较

图 5 提供了我们的方法与两个强基线方法 NAP-ICA [18] 和 SINGAPO [23] 之间的定性比较。每个示例包括：(1) 输入的双状态图像对（闭合和打开），(2) 预测的铰接图，(3) 静止状态下重建的部件布局和关节，以及 (4) 最终的活动状态几何形状。这些示例涵盖了广泛的场景，包括来自 PM 和 ACD 数据集的合成数据。此外，最后三行是真实世界的示例：我们或者从互联网收集静止状态的图像，或者直接拍摄附近物体在两种状态下的图像对。这为评估模型在现有数据集之外的泛化能力提供了一个更真实的场景。对于仅提供静止状态图像的互联网收集示例，我们使用 GPT-4o 来生成相应的活动状态图像，展示了我们方法的灵活性。

与基线方法相比，我们的方法 DIPO 展示了更优的视觉质量和更高的铰接图预测准确率。得益于 PM-X 数据集提供的大规模、结构多样的训练，我们的方法在处理复杂物体或真实世界数据时表现出更好的鲁棒性。此外，在部件密集排列且纹理高度相似的情况下，单图像基线方法常常会混淆，导致错误的铰接推断。相比之下，我们的方法利用静止状态和活动状态之间的对比线索，更准确地识别部件边界、关节连接性和部件运动。

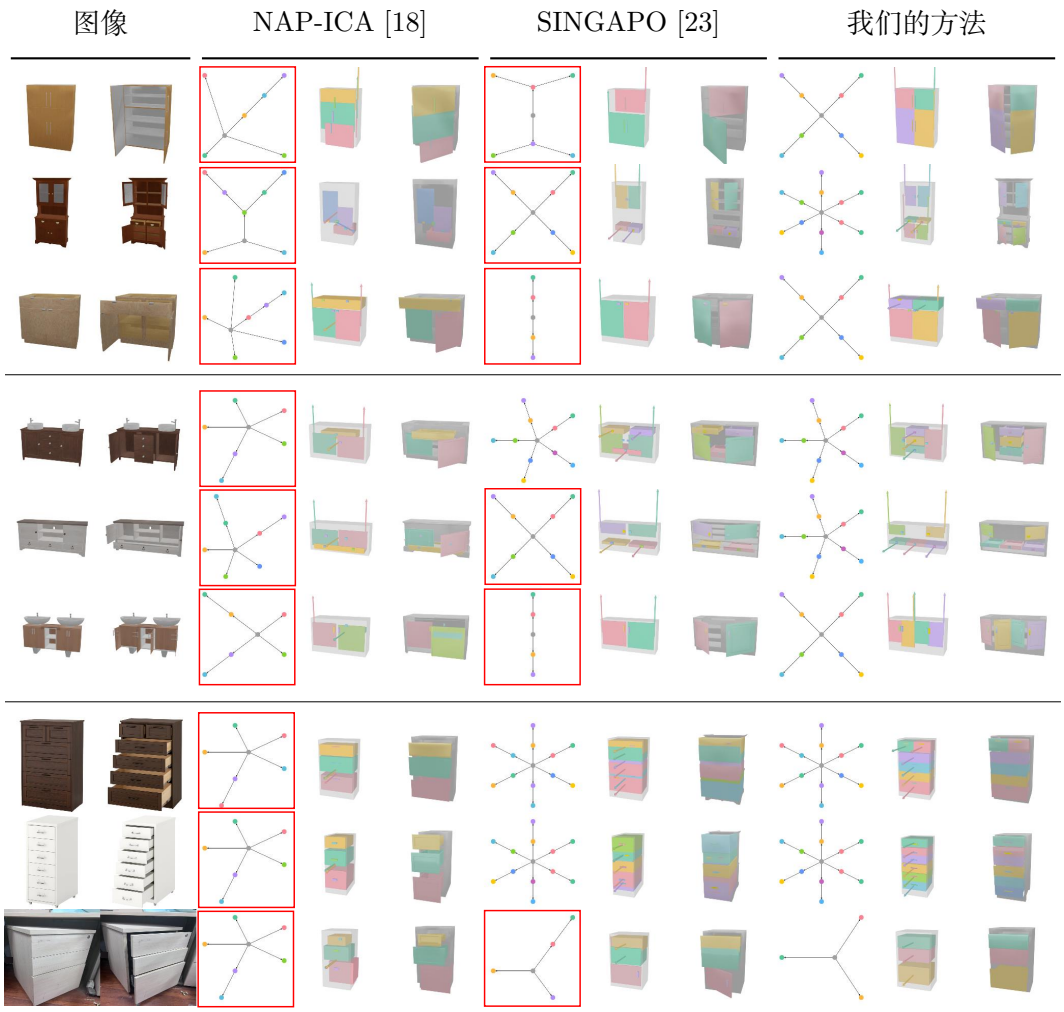


图 5: 所提出的 DIPO 与两个基线方法的视觉比较。前两列展示了双状态图像对。图中还展示了铰接图的预测结果、静止状态下的部件布局和关节可视化，以及活动状态下的最终几何形状。前三行样本来自 PM 数据集，中间三行来自 ACD 数据集，最后三行是真实世界图像。不正确的部件连接用红色框标出。

这些定性结果有力地支持了所提出的 DIPO 的有效性和泛化能力。

5.3 消融实验

我们进行了详细的消融研究，以验证我们框架中每个关键组件的有效性，包括 PM-X 数据集、双状态注入模块 (DIM) 和图推理器 (GR)。我们通过选择性地改变这些组件，构建了几个变体。定量结果总结在表 5 中。此外，我们在以下段落中进一步独立分析了每个组件的设置。

PM-X 数据集的影响。 表 5 显示，在各种消融实验设置中，加入 PM-X 数据集都能持续提高重建质量，表明其具有广泛的有效性。为了进一步验证这一效果，我们额外进行了

表 5: 在 **ACD** 测试集上重建质量和图预测准确率的消融结果。除了 Acc% (↑) 外，其他指标越低越好 (↓)。

设置			重建质量					
PM-X	DIM	GR	RS-d_{gIoU} ↓	AS-d_{gIoU} ↓	RS-d_{cDist} ↓	AS-d_{cDist} ↓	RS-d_{CD} ↓	AS-d_{CD} ↓
			0.9872	0.9900	0.1608	0.2096	0.1083	0.1792
✓			0.9429	0.9464	0.1389	0.1868	0.0849	0.1538
	✓		0.9565	0.9589	0.1478	0.1819	0.0924	0.1407
		✓	0.9902	0.9931	0.1697	0.2157	0.1208	0.1881
✓	✓		0.9212	0.9233	0.1257	0.1589	0.0752	0.1200
✓		✓	0.9332	0.9368	0.1391	0.1843	0.0844	0.1439
	✓	✓	0.9497	0.9515	0.1500	0.1786	0.0973	0.1317
✓	✓	✓	0.9126	0.9151	0.1253	0.1541	0.0751	0.1085

仅使用 25% 和 50% PM-X 数据的实验。如图 6 所示，随着 PM-X 数据比例的增加，静止状态和活动状态的 IoU 分数都稳步下降，证实了 PM-X 在增强结构准确性和泛化能力方面的重要性。

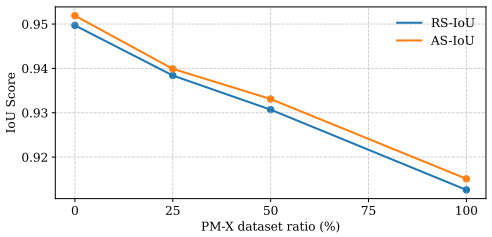


图 6: 不同 PM-X 数据比例下的消融比较。

双图像输入的有效性。 我们进行了消融实验来评估 DIM 模块的贡献。如表 5 所示，添加 DIM 显著改善了所有重建指标的性能。DIM 的有效性进一步体现在图 1 和图 5 中，我们的方法根据活动状态图像准确地识别了运动方向。这表明双图像设计不仅增强了铰接预测，还赋予了模型结构推理的能力。

图推理器分析 如表 5 所示，GR 模块并不能在所有设置下都持续提升性能。这是因为虽然 GR 能够实现更准确的预测，但它也倾向于生成更复杂的拓扑结构。对于没有在 PM-X 数据集上训练的模型变体，这种复杂的图可能成为分布外数据，导致次优性能。然而，当模型使用结构多样的 PM-X 数据集进行训练时，GR 的优势变得更加明显。此外，我们进行了更详细的消融实验来验证 GR 各个组件的有效性。预测准确率的结果见表 4。

表 4: 图推理器的消融结果。

设置	**Acc% ↑**
不含 CoT	39.26
不含视觉输入	37.77
不含双状态输入	39.63
完整模型 (GR)	48.15

6 结论

我们提出了 DIPO，这是一个在处理具有挑战性的数据时，能够提升视觉条件下的铰接物体生成能力的框架。我们设计了一个以静止和活动状态图像对为条件的扩散模型，用于生成铰接式 3D 物体。这种设计提供了更丰富的部件运动信息，从而提高了重建精

度。我们进一步引入了一个基于思维链的图推理器，以增强部件连接关系的预测能力。此外，我们开发了 LEGO-Art，这是一个用于构建多样化且复杂的铰接物体的自动化流水线，并贡献了由该流水线构建的大规模数据集 PM-X。在 PM-X 数据集的支持下，我们的模型实现了更优越的性能和更强的泛化能力。大量的实验验证了我们框架中每个组件的有效性，以及我们的方法相较于现有方法的整体优势。

Acknowledgments and Disclosure of Funding

深圳市科技计划项目 (JCYJ20240813114237048)，“科技甬江 2035”重点技术攻关计划项目 (2024Z120)，中央引导地方科技发展资金项目（科技成果转移转化项目）(254Z0102G)

References

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

[2] Harry G Barrow, Jay M Tenenbaum, Robert C Bolles, and Helen C Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. In *Proceedings: Image Understanding Workshop*, pages 21–27. Science Applications, Inc, 1977.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

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

[5] Xu Chen, Tianjian Jiang, Jie Song, Max Rietmann, Andreas Geiger, Michael J Black, and Otmar Hilliges. Fast-snarf: A fast deformer for articulated neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11796–11809, 2023.

[6] Zoey Chen, Aaron Walsman, Marius Memmel, Kaichun Mo, Alex Fang, Karthikeya Vemuri, Alan Wu, Dieter Fox, and Abhishek Gupta. Urdformer: A pipeline for constructing articulated simulation environments from real-world images. *arXiv preprint arXiv:2405.11656*, 2024.

[7] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21126–21136, 2022.

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

[9] Samir Yitzhak Gadre, Kiana Ehsani, and Shuran Song. Act the part: Learning interaction strategies for articulated object part discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15752–15761, 2021.

[10] Haoran Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, and He Wang. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7081–7091, 2023.

[11] Karol Hausman, Scott Niekum, Sarah Osentoski, and Gaurav S Sukhatme. Active articulation model estimation through interactive perception. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3305–3312. IEEE, 2015.

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

[13] Ruizhen Hu, Wenchao Li, Oliver Van Kaick, Ariel Shamir, Hao Zhang, and Hui Huang. Learning to predict part mobility from a single static snapshot. *ACM Transactions On Graphics (TOG)*, 36(6):1–13, 2017.

[14] Denys Iliash, Hanxiao Jiang, Yiming Zhang, Manolis Savva, and Angel X Chang. S2o: Static to openable enhancement for articulated 3d objects. *arXiv preprint arXiv:2409.18896*, 2024.

[15] Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X Chang, and Manolis Savva. Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16384–16393, 2024.

[16] Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. Parahome: Parameterizing everyday home activities towards 3d generative modeling of human-object interactions. *arXiv preprint arXiv:2401.10232*, 2024.

[17] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[18] Jiahui Lei, Congyue Deng, William B Shen, Leonidas J Guibas, and Kostas Daniilidis. Nap: Neural 3d articulated object prior. *Advances in Neural Information Processing Systems*, 36: 31878–31894, 2023.

[19] Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, et al. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. *arXiv preprint arXiv:2108.03272*, 2021.

[20] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Conference on Robot Learning*, pages 80–93. PMLR, 2023.

[21] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19711–19722, 2024.

[22] Jiayi Liu, Ali Mahdavi-Amiri, and Manolis Savva. Paris: Part-level reconstruction and motion analysis for articulated objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 352–363, 2023.

[23] Jiayi Liu, Denys Iliash, Angel X Chang, Manolis Savva, and Ali Mahdavi-Amiri. Singapo: Single image controlled generation of articulated parts in objects. *arXiv preprint arXiv:2410.16499*, 2024.

[24] Jiayi Liu, Hou In Ivan Tam, Ali Mahdavi-Amiri, and Manolis Savva. Cage: controllable articulation generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17880–17889, 2024.

[25] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024.

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[27] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 909–918, 2019.

[28] Kaichun Mo, Leonidas J Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6813–6823, 2021.

[29] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[30] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[31] Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, et al. Habitat 3.0: A co-habitat for humans, avatars and robots. *arXiv preprint arXiv:2310.13724*, 2023.

[32] Shengyi Qian and David F Fouhey. Understanding 3d object interaction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21753–21763, 2023.

[33] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5020–5030, 2024.

[34] Morgan Quigley, Brian Gerkey, and William D Smart. *Programming Robots with ROS: a practical introduction to the Robot Operating System.* " O'Reilly Media, Inc.", 2015.

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

[36] Bokui Shen, Fei Xia, Chengshu Li, Roberto Martín-Martín, Linxi Fan, Guanzhi Wang, Claudia Pérez-D' Arpino, Shyamal Buch, Sanjana Srivastava, Lyne Tchapmi, et al. igibson 1.0: A simulation environment for interactive tasks in large realistic scenes. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7520–7527. IEEE, 2021.

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[38] Trimble Inc. 3d warehouse, 2025. Accessed: 2025-05-14.

[39] Wei-Cheng Tseng, Hung-Ju Liao, Lin Yen-Chen, and Min Sun. Cla-nerf: Category-level articulated neural radiance field. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 8454–8460. IEEE, 2022.

[40] Xiaogang Wang, Bin Zhou, Yahao Shi, Xiaowu Chen, Qinping Zhao, and Kai Xu. Shape2motion: Joint analysis of motion parts and attributes from 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8876–8884, 2019.

[41] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[42] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Photo wake-up: 3d character animation from a single photo. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5908–5917, 2019.

[43] Yijia Weng, Bowen Wen, Jonathan Tremblay, Valts Blukis, Dieter Fox, Leonidas Guibas, and Stan Birchfield. Neural implicit representation for building digital twins of unknown articulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3141–3150, 2024.

[44] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR, 2022.

[45] Ruiqi Wu, Liangyu Chen, Tong Yang, Chunle Guo, Chongyi Li, and Xiangyu Zhang. Lamp: Learn a motion pattern for few-shot video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7089–7098, 2024.

[46] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11097–11107, 2020.

[47] Zihao Yan, Ruizhen Hu, Xingguang Yan, Luanmin Chen, Oliver Van Kaick, Hao Zhang, and Hui Huang. Rpm-net: recurrent prediction of motion and parts from point cloud. *arXiv preprint arXiv:2006.14865*, 2020.

[48] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul 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.

[49] Ji Yang, Xinxin Zuo, Sen Wang, Zhenbo Yu, Xingyu Li, Bingbing Ni, Minglun Gong, and Li Cheng. Object wake-up: 3d object rigging from a single image. In *European Conference on Computer Vision*, pages 311–327. Springer, 2022.

[50] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

[51] Xuying Zhang, Yupeng Zhou, Kai Wang, Yikai Wang, Zhen Li, Shaohui Jiao, Daquan Zhou, Qibin Hou, and Ming-Ming Cheng. Ar-1-to-3: Single image to consistent 3d object generation via next-view prediction. *arXiv preprint arXiv:2503.12929*, 2025.