PWM-ArtGen: Part World Model for Articulated Object Generation

Wentao Zheng and Ancong Wu*

School of Computer Science and Engineering, Sun Yat-sen University, China
zhengwt28@mail2.sysu.edu.cn, wuanc@mail.sysu.edu.cn

Abstract. The key challenge in articulated 3D object generation from a single image is accurately predicting the underlying kinematic structure. Existing methods either infer kinematic parameters directly from a static image that lacks dynamic part-level kinematic relationships, or estimate parameters from visual dynamics generated from a single image, which is prone to accumulated errors of two steps. Moreover, the limited scale and diversity of existing annotated datasets further hinder generalization to complex, real-world objects. To overcome these limitations, we propose to learn the joint distribution of visual dynamics and kinematic parameters. Recognizing that articulated objects can be formulated as dynamic systems, we propose a unified Part World Model called **PWM-ArtGen**. To leverage unannotated data, this model couples action diffusion and image diffusion with independent diffusion timesteps, which enables visual branch co-training. We further curate a photorealistic dataset of 19.7k part-level image pairs without kinematic annotations, to support co-training. Experiments demonstrate that PWM-ArtGen substantially outperforms existing baselines in the resting state and exhibits strong zero-shot generalization to out-of-distribution objects. We will release our code at https://github.com/Wentap123/PWM-ArtGen.

Keywords: articulated object · visual dynamics · kinematic parameters

1 Introduction

Articulated household objects are ubiquitous in daily environments and serve as fundamental building blocks for interactive virtual worlds, robotics, and embodied AI [33, 43]. Automatically creating high-quality articulated assets at scale remains challenging: expert authoring is costly, and data capture in real settings is often constrained. As a result, increasing attention has been devoted to developing automatic methods for articulated object modeling.

Existing methods fall into two camps: one line of work focuses on optimization-based reconstruction, assumes multi-image or video inputs [23, 26, 27, 32] are available or synthesizes multi-images or video [20, 21, 28] from a single input image. Such approaches inject motion cues and connectivity information to recover articulated behavior, but they often rely on heavy optimization. They also

* Corresponding author.

require careful cross-view or cross-state alignment, or explicit capture of the object in motion, which is not always feasible in practice and limits throughput and scalability to diverse real-world objects. Other approaches [22, 24] regress kinematic parameters directly using trainable feedforward networks trained on existing datasets. While they can produce plausible geometry and kinematics from a resting-state image, they exhibit clear performance degradation when applied to structurally complex or visually ambiguous objects. Since a single static image provides no explicit articulation cues, the absence of dynamic information fundamentally limits identifiability.

Beyond methodological constraints, progress is bottlenecked by the scarcity of high-quality training data. Synthetic datasets like PartNet-Mobility (PM) [29, 44] lack photorealism, often leading to poor real-world generalization. Conversely, realistic datasets like Articulated Containers Dataset (ACD) [8, 15] are too small for large-scale training and are better suited for evaluation. Since acquiring kinematic annotations for real-world objects is prohibitively expensive, exploring methods that can leverage unannotated visual data for training is of great significance. For such unannotated data, we can harness off-the-shelf image and video foundation models to generate or enhance visual priors, easily scaling up diverse, high-quality datasets without relying on costly manual labels.

Motivated by the challenges of single-image articulated generation on real data, we argue that prior work [22, 24] largely neglects part-level appearance dynamics and instead focuses on kinematic parameters. Our key insight is that visual dynamics and kinematics are inherently inter-dependent: observed motions constrain feasible joint types and axes, while kinematic parameters explain motion patterns. We therefore introduce **PWM-ArtGen**, a generative model that learns the joint distribution of part-level visual dynamics and kinematic parameters. By decoupling diffusion timesteps, our approach facilitates visual-branch co-training on unannotated data, thereby capturing complex visual dynamics without the need for manual labels.

Also, we introduce a Visual Dynamics Regularizer (VDR) that explicitly promotes faithful part-level dynamics, thereby limiting dependence on static cues present in a single image. To support co-training, we curate PartNet-Mobility-Reality (PM-R), a photorealistic dataset of 19.7k paired images. Concretely, we apply a frozen image-to-image editing model to PartNet-Mobility renders to enhance materials, lighting, and backgrounds. In summary, our main contributions are as follows:

- **PWM-ArtGen:** We propose a single-image part world model that jointly learns visual dynamics and kinematics, enabling controllable 3D articulated generation and seamless co-training on unannotated data.
- **Visual Dynamics Regularizer (VDR):** We introduce a regularization mechanism that enforces consistency between visual motion cues and articulated behaviors, significantly improving kinematic reasoning.
- **PM-R Dataset:** We construct a photorealistic dataset of 19.7k image pairs with part-level dynamics, which unlocks scalable co-training and effectively bridges the synthetic-to-real gap.

2 Related Work

2.1 Articulated object creation

Reconstruction-based Methods. To advance robotic simulation and digital twins, the rapid generation of articulated object assets from images has garnered significant interest. Forming the foundation of this domain, initial works [8, 9, 13, 15, 29, 38, 44, 45] largely depended on manual annotation to construct extensive datasets. One prominent line of work recovers articulated behavior through per-object optimization. Early approaches in this domain learn deformable signed distance fields [30] or reconstruct parts from point clouds captured across multiple object states [16, 25, 37]. To inject essential motion cues and connectivity information, subsequent methods rely heavily on multi-view imagery, stereo captures [11], or video inputs [23, 26, 27, 32, 34, 39]. Recent advancements even attempt to synthesize these multi-view or video inputs from a single image to guide the optimization process [20, 21, 28]. However, these approaches consistently suffer from low optimization speed, fixed-category assumptions, and a strict dependence on dense cross-view or cross-state alignment. This limits their practical throughput and scalability to diverse, in-the-wild objects.
Generation-based Methods. Another line of work [7,18,22,24] synthesizes articulated objects from a single image by assembling or adapting parts retrieved from pre-constructed asset collections. Although such retrieval-based paradigms benefit from realistic part priors, the absence of dynamic cues fundamentally limits their ability to infer complex real-world articulations. Recent work [42] begins to incorporate motion-related observations, but it relies on dual-image input and therefore cannot be directly applied to the more challenging single-image setting. Beyond retrieval-based generation, several methods [19, 36] attempt to synthesize articulated geometry through implicit representations such as generated SDFs followed by surface reconstruction. However, the reconstruction process can introduce artifacts and often struggles to preserve fine-grained geometric details. Recent advances in 3D foundation models have further enabled generative articulated object modeling [5,6], yet existing solutions are still primarily demonstrated on relatively simple articulation patterns.

Unlike prior single-image methods that focus solely on static geometry or direct parameter regression and thereby overlook dynamic information, we argue that visual dynamics provide crucial physical constraints for kinematic reasoning. PWM-ArtGen jointly models visual dynamics and kinematic parameters. This coupled formulation not only improves parameter estimation but also enables effective co-training on large-scale unannotated data, making our method scalable and generalizable to complex real-world objects.

2.2 World Model

Recent advancements in world models [1–4] have demonstrated remarkable success in learning general physical dynamics from large-scale unannotated videos, which significantly enhances the reasoning and prediction of actions. Building

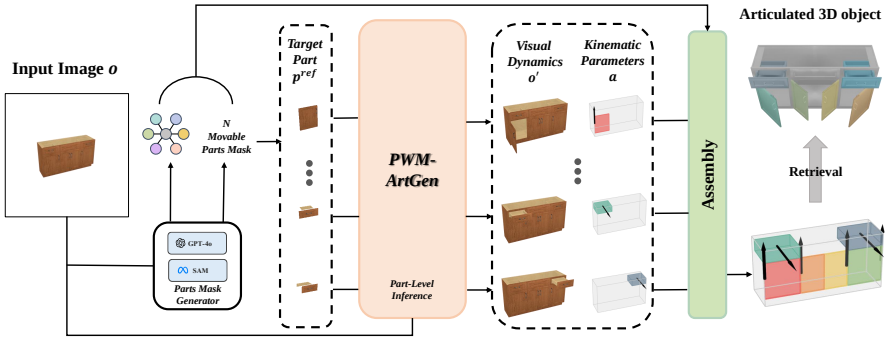

Fig. 1: Overview of PWM-ArtGen, an end-to-end pipeline to convert a single image into an articulated 3D object: (1) **Parts mask generator**: from image o, a pretrained module based on SAM [17] and GPT-4o [14], extracts N part masks $\{m_i\}_{i=1}^{N}$ and an articulate graph for assembly, (2) **Part-level inference**: p_i^{ref} is instantiated as the part mask m_i, sample $(o_i', a_i) \sim p(o', a \mid o, p_i^{ref})$, (3) **Assembly**: align and assemble $\{(o_i', a_i)\}_{i=1}^{N}$ with the shared base bounding box $\mathbf{b}^{base}$ in the object, (4) **Retrieval**: following prior work [22,24] to regularize geometry and motion. Implementation details of the *Part Mask Generator*, *Assembly*, and *Retrieval* are provided in the Supplementary Material.

on this, recent approaches [10, 41, 47] have effectively harnessed video representations to facilitate the prediction of complex robotic trajectories and control actions. Despite these successes, the introduction of world models into the domain of articulated object generation remains largely unexplored, yet it presents a highly promising avenue for capturing intricate structural and kinematic priors. Furthermore, while existing frameworks, *i.e.* UWM [47], are inherently designed for control policies and cannot be directly applied to this generation task, we adapt their core principles to seamlessly fit the unique requirements of articulated object generation.

3 Part World Model

3.1 Overview

We propose a diffusion-based Part World Model that learns the joint distribution of kinematic attributes and dynamic visual representations conditioned on the input image and target articulated part. Drawing on UWM [47], we utilize independent diffusion timesteps to ingest action-free observations. This strategy allows the visual branch to extract kinematically relevant features from unannotated data, significantly enhancing kinematic estimation accuracy. We adopt an articulation-centric parameterization and introduce a visual dynamics regularizer that explicitly promotes faithful part-level dynamics.

Formally, the model learns the conditional distribution $p(o', a \mid o,\ p^{ref})$ through a coupled noise prediction network s_θ, where o denotes the static observation, o' the post-action dynamic observation, a the kinematic attributes, and p^{ref} the target articulated part serving as structural guidance. For each input image, the model generates coherent part-level dynamics and motion parameters that are spatially organized according to p^{ref}, enabling consistent reconstruction of articulated objects. Based on the model, we construct a pipeline that assembles parts into an object generation by retrieval, which is illustrated in Fig. 1.

3.2 Data Representation

We represent each articulated part with a dynamic visual, a set of kinematic attributes, and corresponding part-level structural guidance.

Dynamic visual. The post-action dynamic observation o' is a 256×256 RGB image depicting the scene after executing the action on the reference part.

Kinematic attributes. Compared to prior work [22, 24], we expand the kinematic attributes with $a = \{t, r, \mathbf{l}, \mathbf{d}, \mathbf{b}, \mathbf{b}^{base}\}$, where t denotes the joint type, r the motion range, $\mathbf{l} \in \mathbb{R}^3$ the joint axis location, $\mathbf{d} \in \mathbb{R}^3$ the joint axis direction, $\mathbf{b} \in \mathbb{R}^6$ the part bounding box, and $\mathbf{b}^{base} \in \mathbb{R}^6$ the base bounding box. Considering dynamic visual behavior, the articulation types addressed in this work include revolute and prismatic joints. All articulated parts of the same object share the base bounding box $\mathbf{b}^{base}$, which provides a consistent reference for post-assembly in our part-level modeling framework. This design eliminates the need to predefine a maximum number of parts, as required by prior works.

Part-level structural guidance. The target articulated part p^{ref} is derived from a part mask and comprises both the cropped part image and its 2D geometric cues:

$$p^{\text{ref}} = \{o^{\text{part}},\ 2\text{Dbbox}^{\text{part}},\ 2\text{Dbbox}^{\text{base}}\}. \tag{1}$$

o^{part} means the cropped part observation by the part mask, and $2Dbbox^{part}$, $2Dbbox^{base}$ encode the spatial extent of the region within the image using its center coordinates (c_x, c_y), height h, and width w. We adapt DINOv2 [31] to encode the static observation o into global visual features f_o, and the cropped part observation o^{part} into local features $f_{o^{part}}$. To ensure scale invariance, $2Dbbox^{part}$, $2Dbbox^{base}$ are normalized as $f_{2Dbbox^{part}}$, $f_{2Dbbox^{base}}$ by the corresponding resolution of the original image. The resulting reference part embedding f^{ref} is defined as:

$$f^{\text{ref}} = \{f_{o^{\text{part}}},\ f_{2\text{Dbbox}^{\text{part}}},\ f_{2\text{Dbbox}^{\text{base}}}\}. \tag{2}$$

Consequently, a training sample is parameterized by the tuple (o', a, o, p^{ref}), which simplifies to $(o', \emptyset, o, p^{ref})$ during the co-training phase to incorporate unannotated visual dynamics.

3.3 Model Architecture

We instantiate PWM-ArtGen as a diffusion Transformer that jointly models dynamic visual tokens and kinematic attributes, as shown in Fig. 2. The model

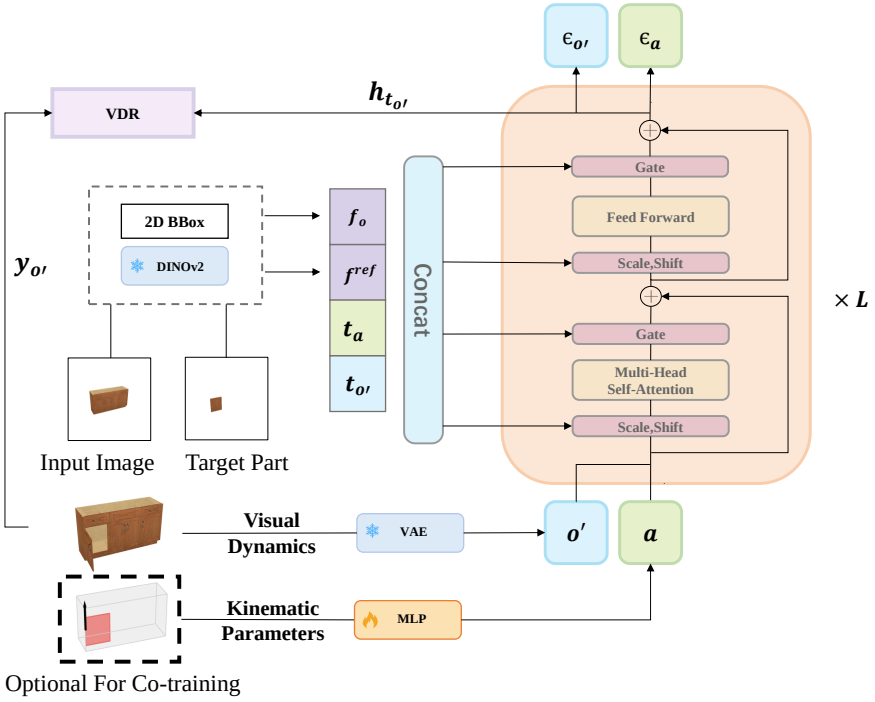

Fig. 2: A single Part World Model (PWM-ArtGen) block consists of a transformer block with observation, a target articulated part, and diffusion timesteps conditioning via adaptive layer norm.

predicts the actions and observation noises ϵ_a and $\epsilon_{o'}$, conditioned on static observations o and the reference part p^{ref}, where o is the input image and p^{ref} specifies the target articulated part. In accordance with the representation established in Sec. 3.2, the static observation o is encoded as global visual features f_o, and the reference part p^{ref} is embedded as f^{ref}. The concatenation of f^{ref} with independent diffusion timesteps t_a and $t_{o'}$ serves as the conditioning signal for the transformer blocks via adaptive layer normalization (AdaLN). For the visual branch, we operate in a latent image space by a frozen SDXL VAE and patchify the latent into N tokens of width D, yielding $o' \in \mathbb{R}^{N \times D}$. Concurrently, for the kinematic branch, the kinematic attributes are projected into a tokenized space via a trainable MLP, denoted as $a \in \mathbb{R}^{20 \times D}$ and concatenated with the visual tokens to form the input.

3.4 Visual Dynamics Regularizer

Inspired by REPA [46], we introduce a visual dynamics regularizer (VDR) to promote faithful part-level dynamics and mitigate reliance on static single-image evidence. Under joint training of action and visual branches, VDR biases the

visual representations toward dynamics-consistent features that reflect action-induced motion. During training, the diffusion model produces joint hidden states $h_{t_{o'}}, h_{t_a} = f_\theta(z_t)$ at timestep t, where $z_t = (a_{t_a}, o'_{t_{o'}})$ and $h^{[n]}_{t_{o'}}$ denotes the feature of the n-th spatial patch in the visual branch. To ensure semantic consistency between the generated visual patches and the pretrained visual encoder, we apply the visual dynamics regularizer loss defined as

$$\mathcal{L}_{\mathrm{VDR}}(\theta) := -\,\mathbb{E}_{o', \epsilon_{o'}, t_{o'}, t_a}\left[\frac{1}{N}\sum_{n=1}^{N} \mathrm{sim}\left(y^{[n]}_{o'},\, h^{[n]}_{t_{o'}}\right)\right], \tag{3}$$

where $y^{[n]}_{o'} = f(x_{o'})$ denotes the n-th patch feature of the clean image, and $\mathrm{sim}(\cdot,\cdot)$ is the cosine similarity function. This aligns the denoising trajectory with the semantic manifold of clean data, promoting robustness to noise while preserving essential visual structure. The network in Fig. 2 consists of L layers in total, and only the visual hidden states from the first l layers are aligned with the DINOv2 patch-wise features of the post-action observation o' as shown in Eq. (3).

3.5 Co-Training

PWM-ArtGen learns a joint diffusion process over kinematic parameters a and post-action dynamic observation o' conditioned on static observation o and target articulated part p^{ref}, enabling the model to incorporate part-level structural guidance during both training and inference. Specifically, we train a coupled noise prediction network [12] $s_\theta(o, p^{\mathrm{ref}}, a_{t_a}, o'_{t_{o'}}, t_a, t_{o'})$ that approximates the conditional expectation of action and observation noise.

To leverage both action-annotated and action-free data, we adopt a co-training strategy. Let $\mathcal{D}_{\mathrm{paired}}$ denote the dataset with ground-truth actions and $\mathcal{D}_{\mathrm{free}}$ denote the dataset containing only observations. For action-free samples, we impute the missing actions with pure Gaussian noise by fixing the action diffusion timestep to T (*i.e.*, $a_T = \epsilon_a \sim \mathcal{N}(0,1)$). The corresponding denoising objective is formulated as follows:

$$\begin{aligned}\mathcal{L}_{\epsilon_a,\,\epsilon_{o'}}(\theta) = &\mathbb{E}_{\substack{(o,\,p^{\mathrm{ref}},\,a,\,o')\sim\mathcal{D}_{\mathrm{paired}}\\ t_a,\,t_{o'}\sim\mathcal{U}(0,T)\\ \epsilon_a,\,\epsilon_{o'}\sim\mathcal{N}(0,1)}}\Big[w_a\|\epsilon^\theta_a - \epsilon_a\|_2^2 + w_{o'}\|\epsilon^\theta_{o'} - \epsilon_{o'}\|_2^2\Big]\\ &+\gamma\mathbb{E}_{\substack{(o,\,p^{\mathrm{ref}},\,\emptyset,\,o')\sim\mathcal{D}_{\mathrm{free}}\\ t_a=T,\,t_{o'}\sim\mathcal{U}(0,T)\\ \epsilon_a,\,\epsilon_{o'}\sim\mathcal{N}(0,1)}}\Big[w_a\|\epsilon^\theta_a - \epsilon_a\|_2^2 + w_{o'}\|\epsilon^\theta_{o'} - \epsilon_{o'}\|_2^2\Big],\end{aligned} \tag{4}$$

where

$$\epsilon^\theta_a,\, \epsilon^\theta_{o'} = s_\theta(o,\, p^{\mathrm{ref}},\, a_{t_a},\, o'_{t_{o'}},\, t_a,\, t_{o'}), \tag{5}$$

$$a_{t_a} = \sqrt{\bar\alpha_{t_a}}\, a + \sqrt{1-\bar\alpha_{t_a}}\, \epsilon_a, \tag{6}$$

$$o'_{t_{o'}} = \sqrt{\bar\alpha_{t_{o'}}}\, o' + \sqrt{1-\bar\alpha_{t_{o'}}}\, \epsilon_{o'}. \tag{7}$$

Here, w_a and $w_{o'}$ are weights chosen to trade off between the action prediction and image prediction objectives, and γ is a hyperparameter that balances the contribution of the action-free data. For any time step $t \in \{1, \ldots, T\}$, we define $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, where $\{\beta_s\}_{s=1}^{T}$ is a fixed variance schedule as in DDPM [12]. The timesteps t_a and $t_{o'}$ are sampled independently, meaning $\bar{\alpha}_{t_a}$ and $\bar{\alpha}_{t_{o'}}$ correspond to distinct noise levels for the action and observation diffusion processes, respectively. This conditioning allows our model to align motion dynamics and visual appearance with part-level spatial context, providing a structured prior for generating coherent articulated behaviors.

The final training objective combines the co-training loss with the visual dynamics regularization (VDR):

$$\mathcal{L} = \mathcal{L}_{\epsilon_a, \epsilon_{o'}} + \lambda \mathcal{L}_{\text{VDR}}, \tag{8}$$

where λ balances VDR loss with diffusion reconstruction accuracy. This joint training encourages articulation-aware dynamics while maintaining semantic coherence in the generated visual space.

3.6 Synthetic-Real Mixed Training Strategy

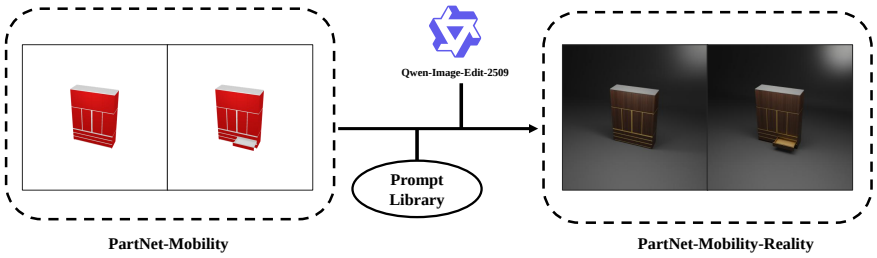

Fig. 3: Construction pipeline of PartNet-Mobility-Reality (PM-R). Synthetic renderings from PartNet-Mobility are enhanced into photorealistic counterparts through the Qwen-Image-Edit-2509 [40] model guided by a structured Prompt Library, producing 19.7k photorealistic samples.

To enhance the model's generalization ability on real-world images, we construct the PM-R dataset as illustrated in Fig. 3. During co-training, we employ a synthetic–real mixed strategy using a length-proportional interleaving sampler that visits all samples once per epoch while mixing the two domains according to their relative sizes. At each epoch, the PartNet-Mobility-Reality and PartNet-Mobility subsets are independently shuffled, and a weighted round-robin procedure generates a global index order that maintains a consistent real–synthetic ratio. The sampler is DDP-compatible and partitions the sequence across workers without overlap, providing iteration-level balanced mixing. During training, the model jointly learns to predict both the action a and the post-action observation o', enabling a unified understanding of motion and visual dynamics.

4 Experiments

We evaluate reconstruction quality and part connectivity graph prediction on both the ACD and PartNet-Mobility test sets. Qualitative results on ACD are conditioned on the ground-truth part graph. Furthermore, we comprehensively evaluate the visual dynamics of the image branch through both quantitative metrics and qualitative visualizations on the PartNet-Mobility test set. Ablation studies validate the contributions of key components: the image branch (IB) for joint modeling, the Visual Dynamics Regularizer (VDR), and co-training (Co-T) on PartNet-Mobility-Reality (PM-R) dataset. Additional qualitative results and experiments are provided in the Supplementary Material.

4.1 Experimental Setup

Datasets. Following [22], we collect data from the PartNet-Mobility [29] dataset to train our model, and from the ACD dataset [15] for additional evaluation. Our training set consists of 473 articulated objects (985 movable parts in total) from PartNet-Mobility. Each object is rendered using `BLENDER_EEVEE_NEXT` under both the rest and open states, with 20 images per state sampled from random viewpoints within a 90° horizontal and 90° upward range relative to the object's front. We extract part-level masks for the rest state and apply realistic augmentation strategies to construct PM-R. In total, we obtain about 19.7K rendered samples of PM for training and 19.7k PM-R image pairs without kinematic annotations for co-training. For testing, we use 77 unseen objects (155 parts) from PartNet-Mobility as the test split, each rendered from two random views in both states, yielding 310 test samples. To further evaluate the model's generalization capability on realistic data, we include 135 objects (455 parts) from the ACD dataset in a zero-shot setting, producing 910 additional samples that better reflect real-world scenarios. To evaluate the visual branch, we render ten random views in both states from the PM test split, yielding 1550 test samples. Please refer to the Supplementary Material for more details on rendering configurations.
Compared methods. We select some representative methods: URDFormer [7], NAP [19], SINGAPO [22], and Articulate-Anything [18] for comparison of kinematic parameters; DragAPart [20], and Puppet-Master [21] for comparison of visual dynamics. Due to the limited complexity of parts in the PartNet-Mobility dataset, these methods would suffer significant performance degradation, particularly in zero-shot scenarios, if not trained on larger datasets with more diverse part configurations, as reported in SINGAPO [22]. In contrast, our method is specifically designed for part-level modeling and is thus largely unaffected by such dataset limitations. To ensure a fair comparison, we evaluated SINGAPO using its released checkpoint trained on 3,063 objects (approximately six times larger than PartNet-Mobility) and adopted the results of URDFormer and NAP-ICA reported in [22], where both methods were strictly adapted to the SINGAPO framework based on [7, 19]. Furthermore, we evaluate Articulate-Anything [18] using GPT-4o [14], and directly utilize the officially released checkpoints for the visual dynamics baselines, DragAPart [20] and Puppet-Master [21].

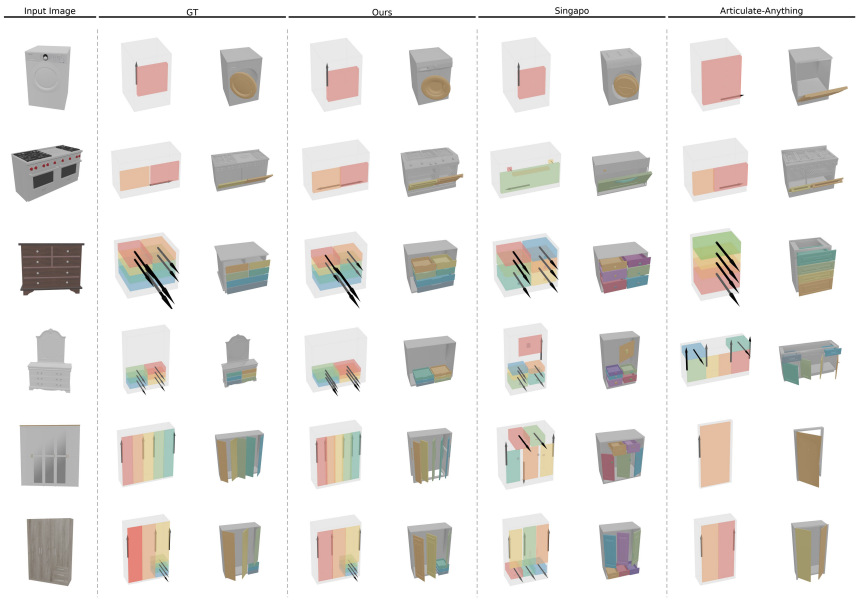

Fig. 4: Qualitative comparison on the ACD dataset under zero-shot evaluation. From left to right: input image, ground truth (GT), our method, SINGAPO, and Articulate-Anything. Each row shows the predicted part layout and motion axes in both resting and articulated states. The last four rows highlight that other methods' predictions exhibit spatial misalignment with GT in part positioning and joint orientation. In contrast, our method produces parts that are **spatially better aligned with GT**, with more accurate joint orientations, cleaner boundaries, and more realistic articulation dynamics—especially on objects with complex textures and occlusions.

Metrics. To evaluate reconstruction quality and articulation correctness, we adopt four quantitative metrics: (1) d_{gIoU} $\downarrow$, the generalized IoU between predicted and ground-truth part bounding boxes; (2) d_{cDist} $\downarrow$, the Euclidean distance between the centers of corresponding parts; (3) d_{CD} $\downarrow$, the Chamfer Distance between predicted and ground-truth meshes; and (4) Acc $\uparrow$, the accuracy of the predicted articulation graph. In addition, we report the average overlapping ratio AOR $\downarrow$ introduced by CAGE [24] to detect unrealistic collisions among sibling parts in the articulated graph. All metrics are computed over both resting and articulated states. For clarity, we prefix the metric names with **RS-**(resting state) and **AS-**(articulated state) in the tables to indicate the corresponding evaluation state. We evaluate all methods under two settings: with and without access to the ground-truth part connectivity graphs and masks. Following [20], we evaluate the generated visual dynamics using standard image quality metrics, including pixel-level PSNR, patch-level SSIM, and feature-level LPIPS. We utilize ground-truth drag or part mask as a conditioning signal.

Table 1: Comparison of reconstruction quality and graph prediction accuracy on the ACD test set.

	Reconstruction quality						Collision	Graph
Method	RS-d_{gIoU} ↓	AS-d_{gIoU} ↓	RS-d_{cDist} ↓	AS-d_{cDist} ↓	RS-d_{CD} ↓	AS-d_{CD} ↓	AOR↓	Acc%↑
URDFormer-GTbbox	1.1986	1.2012	0.2292	0.2931	0.4343	0.4965	0.1209	48.53
NAP-ICA-GTgraph	1.0585	1.0631	0.1987	0.2810	0.1376	0.2417	0.0193	36.67
SINGAPO-GTgraph	0.9729	0.9767	0.1589	0.1968	0.1177	0.1752	0.0112	100.00
Ours-GTgraph	**0.5693**	**0.5790**	**0.0687**	**0.1144**	**0.0569**	**0.0813**	**0.0031**	100.00
URDFormer	1.2288	1.2309	0.2914	0.4285	0.7198	0.8995	0.2840	4.48
NAP-ICA	1.0233	1.0286	0.1691	0.2331	0.1110	0.1887	0.0133	16.67
SINGAPO	0.9775	0.9810	0.1574	0.2016	0.1100	0.1744	0.0103	39.34
Articulate-Anything	0.7574	0.7649	0.1938	0.2684	0.1826	0.2861	**0.0039**	40.00
Ours	**0.6822**	**0.6914**	**0.1480**	**0.1876**	**0.1033**	**0.1385**	0.0906	39.34

Table 2: Comparison of reconstruction quality and graph prediction accuracy on the PartNet-Mobility test set.

	Reconstruction quality						Collision	Graph
Method	RS-d_{gIoU} ↓	AS-d_{gIoU} ↓	RS-d_{cDist} ↓	AS-d_{cDist} ↓	RS-d_{CD} ↓	AS-d_{CD} ↓	AOR↓	Acc%↑
URDFormer-GTbbox	1.0861	1.0882	0.1471	0.3225	0.3400	0.6031	0.0616	83.55
NAP-ICA-GTgraph	0.6830	0.6915	0.0739	0.2206	0.0282	0.2646	0.0105	44.81
SINGAPO-GTgraph	0.4345	0.4506	**0.0319**	**0.0751**	**0.0108**	0.0840	0.0019	100.00
Ours-GTgraph	**0.3741**	**0.3803**	0.0455	0.0913	0.0193	**0.0678**	**0.0003**	100.00
URDFormer	1.1868	1.1879	0.2693	0.4535	0.5502	0.8374	0.2341	32.03
NAP-ICA	0.5778	0.5854	0.0501	0.0979	**0.0173**	**0.0914**	0.0120	75.97
SINGAPO	0.4934	0.5065	**0.0471**	**0.0968**	0.0199	0.1107	0.0033	82.47
Articulate-Anything	0.4731	0.4870	0.1561	0.2302	0.1649	0.3126	**0.0025**	59.74
Ours	**0.4623**	**0.4707**	0.0890	0.1415	0.0446	0.1272	0.0434	82.47

4.2 Implementation Details

We train PWM-ArtGen for 200k steps on 4×RTX 4090 GPUs with a global batch size of 128. The diffusion transformer features 12 layers, 12 attention heads, a hidden dimension of 768, an MLP ratio of 4, 2×2 patch embeddings, and 512-dimensional timestep embeddings. Optimization uses AdamW (learning rate 1×10^{-4}, weight decay 1×10^{-6}, $\beta = (0.9, 0.999)$, $\epsilon = 1 \times 10^{-8}$). Inference relies on 10-step DDIM sampling [12, 35]. For the training objectives, we set the action and image prediction weights to $w_a = w_{o'} = 1.0$, and the action-free data weight to $\gamma = 1.0$. To facilitate early visual-kinematic alignment, we apply the Visual Dynamics Regularizer (VDR) only during the first 80k training steps, with $\lambda = 0.5$, utilizing spatial features from a frozen DINOv2-Reg/14 encoder [31]. Additional details are in the Supplementary Material.

4.3 Comparison with the State-of-the-art Methods

Quantitative Results On Kinematic. We report quantitative results on the ACD and PartNet-Mobility datasets in Tab. 1 and Tab. 2, respectively. For fair comparison, we use the same part connection graph predicted by GPT-4o [14].

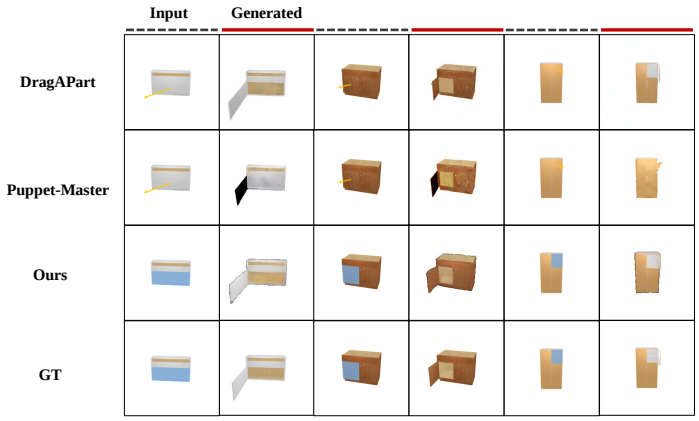

Fig. 5: Qualitative comparison of interactive dynamics generation. Compared to DragAPart and Puppet-Master, which utilize *drag-based* prompts (yellow arrows), our model adopts *mask-based* conditioning (blue regions). Black dashed lines and red solid lines denote the input and generated frames, respectively. Despite the significantly reduced parameter count, our method produces results consistent with the Ground Truth (GT) and competitive with state-of-the-art baselines.

All diffusion-based generative methods are evaluated five times per test sample, and the averaged results are reported to mitigate stochastic variations.

On the ACD test set, which contains more realistic and diverse articulated objects, our method achieves clearly superior performance across all metrics. Compared to existing baselines, PWM-ArtGen delivers the lowest reconstruction errors, demonstrating strong generalization capability to real-world-like data. Such consistent superiority on ACD highlights our model's robustness under domain shift, which is a key challenge for articulated object generation.

Table 3: Quantitative evaluation of the dynamics branch on the PM test set.

Method	PSNR↑	SSIM↑	LPIPS↓	Parameters	Time
DragAPart	**24.803**	**0.943**	**0.079**	1.43 B	6.9 s
Puppet-Master	24.440	0.941	0.084	1.68 B	11.6 s
Ours	24.328	0.906	0.093	**0.38 B**	**0.7 s**

On the PartNet-Mobility test set, when the ground-truth graph and masks are provided, PWM-ArtGen performs on par or better than the strongest baseline across most metrics, with only marginal differences in a few cases. In the setting without ground-truth graph and masks, our method still outperforms competing approaches on several measures, though it exhibits slightly higher d_{cDist} and d_{CD} values. This minor degradation is mainly attributed to the current limitations of our Part Mask Generator, especially on synthetic datasets

such as PM, where rendering artifacts and imperfect geometry annotations often lead to suboptimal mask quality. In contrast, the ACD dataset provides relatively cleaner and more realistic object geometry, resulting in more reliable part-level supervision. It is worth noting that improving the segmentation quality of the Part Mask Generator is not the primary focus of this work, but this limitation can be effectively addressed in future research through more advanced segmentation pipelines. Overall, PWM-ArtGen exhibits strong reconstruction accuracy, reliable articulation prediction, and remarkable generalization to realistic data, which is one of the most challenging aspects of articulated generation.

Qualitative Comparison On Kinematic. As illustrated in Fig. 4, PWM-ArtGen generates more consistent and physically plausible part motions than others. Our method preserves structural integrity during articulation and better aligns part boundaries, demonstrating improved understanding of 3D part relationships. The improvement is especially evident on the ACD dataset, where the objects exhibit complex materials and realistic lighting conditions, highlighting the strong generalization ability of PWM-ArtGen to real-world scenarios.

Comparison On visual. As illustrated in Fig. 5, our model produces qualitative results comparable to Puppet-Master, while trailing slightly behind DragAPart. However, this gap is well-justified by the significant efficiency gains detailed in Tab. 3: our model achieves nearly 10× faster inference speed with only 0.25× the parameters of DragAPart. These results indicate that our dynamics branch can generate visually plausible outputs. Furthermore, subsequent ablation studies demonstrate that explicit dynamics modeling significantly enhances the accuracy of kinematic estimation.

4.4 Ablation Study

We conduct detailed ablation studies to verify the effectiveness of each key component in **PWM-ArtGen**, including the Image Branch (IB) for joint modeling of visual dynamics and kinematics, the Visual Dynamics Regularizer (VDR), and the co-training (Co-T) on the PartNet-Mobility-Reality (PM-R) dataset. We construct several variants by selectively altering these components. The quantitative results are summarized in Tab. 4 and Tab. 5.

Effectiveness of Image Branch. We conduct ablation experiments to assess the contribution of jointly models visual dynamics and kinematic parameters. As shown in Tab. 4, modeling dynamics provides a strong inductive bias for kinematic estimation, leads to a noticeable degradation across all metrics rather than only action branch. Without the guidance of visual cues, the model struggles to infer motion directions and part boundaries accurately.

The Impact of co-training. Ablation results in Tab. 4 and Tab. 5 demonstrate that omitting co-training on PM-R severely degrades performance under complex articulated states. Since PM-R closely aligns with real-world data distributions, its inclusion during co-training effectively bridges the synthetic-to-real gap. It enhances the model's ability to handle near-real inputs from ACD, leading to more precise kinematic estimation.

Table 4: Ablation study on key design components for kinematic parameters. Results are reported on the ACD dataset, assuming a ground truth (GT) part connectivity graph.

Settings			Reconstruction Quality						Collision
IB	VDR	Co-T	RS-d_{gIoU}↓	AS-d_{gIoU}↓	RS-d_{cDist}↓	AS-d_{cDist}↓	RS-d_{CD}↓	AS-d_{CD}↓	AOR↓
			0.6567	0.6684	0.0801	0.1877	0.0796	0.1520	0.0252
✓			0.5948	0.6045	0.0739	0.1202	0.0681	0.0915	0.0033
✓	✓		0.5885	0.5986	0.0699	0.1172	0.0587	0.0840	0.0038
✓		✓	0.5786	0.5878	0.0696	0.1174	0.0619	0.0854	**0.0020**
✓	✓	✓	**0.5693**	**0.5790**	**0.0687**	**0.1144**	**0.0569**	**0.0813**	0.0031

Table 5: Ablation study on key design components for visual dynamics. Results are reported on the PM dataset.

Settings			Visual Dynamics Quality		
IB	VDR	Co-T	PSNR↑	SSIM↑	LPIPS↓
✓			21.964	0.835	0.132
✓	✓		23.383	0.886	0.105
✓		✓	22.908	**0.908**	0.107
✓	✓	✓	**24.328**	0.906	**0.093**

Analysis of Visual Dynamics Regularizer. As shown in Tab. 4 and Tab. 5, removing the VDR compromises part-wise coherence. This module imposes a critical inductive bias that encourages consistent motion and smoother surface deformations, thereby bridging the gap between static geometry and dynamic articulation.

5 Conclusion

In this work, we presented PWM-ArtGen, a unified Diffusion Transformer framework for generating articulated objects from a single image. By jointly modeling part-level visual dynamics and kinematic parameters through decoupled diffusion timesteps, our approach achieves a strong inductive bias for kinematic estimation without predefining part counts. Co-training on the PM-R dataset effectively leverages action-free visual priors and bridges the synthetic-to-real gap, while the Visual Dynamics Regularizer (VDR) ensures part-wise coherence and physically consistent motion. PWM-ArtGen achieves state-of-the-art performance on ACD and competitive results on PartNet-Mobility, demonstrating robust zero-shot generalization.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2023YFA1008501).

References

1. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575 (2025)
2. Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., Ballas, N.: Revisiting feature prediction for learning visual representations from video. arXiv preprint arXiv:2404.08471 (2024)
3. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators (2024), `https://openai.com/research/video-generation-models-as-world-simulators`, accessed: 29 June 2026
4. Bruce, J., Dennis, M., Edwards, A., Parker-Holder, J., Shi, Y.J., Hughes, E., Lai, M., Mavalankar, A., Steigerwald, R., Apps, C., Aytar, Y., Bechtle, S., Behbahani, F., Chan, S., Heess, N., Gonzalez, L., Osindero, S., Ozair, S., Reed, S., Zhang, J., Zolna, K., Clune, J., De Freitas, N., Singh, S., Rocktäschel, T.: Genie: generative interactive environments. In: Int. Conf. Mach. Learn. (2024)
5. Chen, C., Liu, I., Wei, X., Su, H., Liu, M.: Freeart3d: Training-free articulated object generation using 3d diffusion. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers (2025)
6. Chen, H., Lan, Y., Chen, Y., Pan, X.: Artilatent: Realistic articulated 3d object generation via structured latents. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers (2025)
7. Chen, Z., Walsman, A., Memmel, M., Mo, K., Fang, A., Vemuri, K., Wu, A., Fox, D., Gupta, A.: Urdformer: A pipeline for constructing articulated simulation environments from real-world images. In: Proceedings of Robotics: Science and Systems (2024)
8. Collins, J., Goel, S., Deng, K., Luthra, A., Xu, L., Gundogdu, E., Zhang, X., Vicente, T.F.Y., Dideriksen, T., Arora, H., Guillaumin, M., Malik, J.: Abo: Dataset and benchmarks for real-world 3d object understanding. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 21094–21104 (2022)
9. Geng, H., Xu, H., Zhao, C., Xu, C., Yi, L., Huang, S., Wang, H.: Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7081–7091 (2023)
10. Guo, Y., Hu, Y., Zhang, J., Wang, Y.J., Chen, X., Lu, C., Chen, J.: Prediction with action: Visual policy learning via joint denoising process. In: Adv. Neural Inform. Process. Syst. vol. 37, pp. 112386–112410 (2024)
11. Heppert, N., Irshad, M.Z., Zakharov, S., Liu, K., Ambrus, R.A., Bohg, J., Valada, A., Kollar, T.: Carto: Category and joint agnostic reconstruction of articulated objects. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 21201–21210 (2023)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Adv. Neural Inform. Process. Syst. vol. 33, pp. 6840–6851 (2020)

13. Hu, R., Li, W., Van Kaick, O., Shamir, A., Zhang, H., Huang, H.: Learning to predict part mobility from a single static snapshot. ACM Trans. Graph. **36**(6), 1–13 (2017)
14. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
15. Iliash, D., Jiang, H., Zhang, Y., Savva, M., Chang, A.X.: S2o: Static to openable enhancement for articulated 3d objects. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6785–6795 (2026)
16. Jiang, Z., Hsu, C.C., Zhu, Y.: Ditto: Building digital twins of articulated objects from interaction. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5616–5626 (2022)
17. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: Int. Conf. Comput. Vis. (2023)
18. Le, L., Xie, J., Liang, W., Wang, H.J., Yang, Y., Ma, Y.J., Vedder, K., Krishna, A., Jayaraman, D., Eaton, E.: Articulate-anything: Automatic modeling of articulated objects via a vision-language foundation model. In: Int. Conf. Learn. Represent. (2025)
19. Lei, J., Deng, C., Shen, W.B., Guibas, L., Daniilidis, K.: Nap: Neural 3d articulated object prior. In: Adv. Neural Inform. Process. Syst. vol. 36, pp. 31878–31894 (2023)
20. Li, R., Zheng, C., Rupprecht, C., Vedaldi, A.: Dragapart: Learning a part-level motion prior for articulated objects. In: Eur. Conf. Comput. Vis. pp. 165–183. Springer (2024)
21. Li, R., Zheng, C., Rupprecht, C., Vedaldi, A.: Puppet-master: Scaling interactive video generation as a motion prior for part-level dynamics. In: Int. Conf. Comput. Vis. pp. 13405–13415 (2025)
22. Liu, J., Iliash, D., Chang, A., Savva, M., Mahdavi Amiri, A.: Singapo: Single image controlled generation of articulated parts in objects. In: Int. Conf. Learn. Represent. (2025)
23. Liu, J., Mahdavi-Amiri, A., Savva, M.: Paris: Part-level reconstruction and motion analysis for articulated objects. In: Int. Conf. Comput. Vis. pp. 352–363 (2023)
24. Liu, J., Tam, H.I.I., Mahdavi-Amiri, A., Savva, M.: Cage: Controllable articulation generation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 17880–17889 (2024)
25. Liu, S., Gupta, S., Wang, S.: Building rearticulable models for arbitrary 3d objects from 4d point clouds. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 21138–21147 (2023)
26. Liu, Y., Jia, B., Lu, R., Gan, C., Chen, H., Ni, J., Zhu, S.C., Huang, S.: Videoartgs: Building digital twins of articulated objects from monocular video. arXiv preprint arXiv:2509.17647 (2025)
27. Liu, Y., Jia, B., Lu, R., Ni, J., Zhu, S.C., Huang, S.: Building interactable replicas of complex articulated objects via gaussian splatting. In: Int. Conf. Learn. Represent. (2025)
28. Lu, R., Liu, Y., Tang, J., Ni, J., Wang, Y., Wan, D., Zeng, G., Chen, Y., Huang, S.: Dreamart: Generating interactable articulated objects from a single image. arXiv preprint arXiv:2507.05763 (2025)
29. Mo, K., Zhu, S., Chang, A.X., Yi, L., Tripathi, S., Guibas, L.J., Su, H.: Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 909–918 (2019)

30. Mu, J., Qiu, W., Kortylewski, A., Yuille, A., Vasconcelos, N., Wang, X.: A-sdf: Learning disentangled signed distance functions for articulated shape representation. In: Int. Conf. Comput. Vis. pp. 13001–13011 (2021)
31. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. Trans. Mach. Learn Res. (2024)
32. Peng, W., Lv, J., Lu, C., Savva, M.: itaco: Interactable digital twins of articulated objects from casually captured rgbd videos. In: International Conference on 3D Vision (3DV). pp. 520–531 (2026)
33. Shen, B., Xia, F., Li, C., Martín-Martín, R., Fan, L., Wang, G., Pérez-D'Arpino, C., Buch, S., Srivastava, S., Tchapmi, L., et al.: igibson 1.0: A simulation environment for interactive tasks in large realistic scenes. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 7520–7527. IEEE (2021)
34. Song, C., Wei, J., Foo, C.S., Lin, G., Liu, F.: Reacto: Reconstructing articulated objects from a single video. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5384–5395 (2024)
35. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: Int. Conf. Learn. Represent. (2021)
36. Su, J., Feng, Y., Li, Z., Song, J., He, Y., Ren, B., Xu, B.: Artformer: Controllable generation of diverse 3d articulated objects. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1894–1904 (2025)
37. Tseng, W.C., Liao, H.J., Yen-Chen, L., Sun, M.: Cla-nerf: Category-level articulated neural radiance field. In: International Conference on Robotics and Automation. pp. 8454–8460 (2022)
38. Wang, X., Zhou, B., Shi, Y., Chen, X., Zhao, Q., Xu, K.: Shape2motion: Joint analysis of motion parts and attributes from 3d shapes. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 8876–8884 (2019)
39. Weng, Y., Wen, B., Tremblay, J., Blukis, V., Fox, D., Guibas, L., Birchfield, S.: Neural implicit representation for building digital twins of unknown articulated objects. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3141–3150 (2024)
40. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
41. Wu, H., Jing, Y., Cheang, C., Chen, G., Xu, J., Li, X., Liu, M., Li, H., Kong, T.: Unleashing large-scale video generative pre-training for visual robot manipulation. In: Int. Conf. Learn. Represent. (2024)
42. Wu, R., wang, X., Liu.Liu, Guo, C.L., Qiu, J., Li, C., Huang, L., Su, Z., Cheng, M.M.: Dipo: Dual-state images controlled articulated object generation powered by diverse data. In: Adv. Neural Inform. Process. Syst. vol. 38, pp. 108665–108689 (2025)
43. Wu, Y., Lin, Y., Lao, W., Lin, Y., Wei, Y.L., Zheng, W.S., Wu, A.: Dexgrasp-zero: A morphology-aligned policy for zero-shot cross-embodiment dexterous grasping. arXiv preprint arXiv:2603.16806 (2026)
44. Xiang, F., Qin, Y., Mo, K., Xia, Y., Zhu, H., Liu, F., Liu, M., Jiang, H., Yuan, Y., Wang, H., Yi, L., Chang, A.X., Guibas, L.J., Su, H.: Sapien: A simulated part-based interactive environment. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 11094–11104 (2020)
45. Yan, Z., Hu, R., Yan, X., Chen, L., Van Kaick, O., Zhang, H., Huang, H.: Rpm-net: recurrent prediction of motion and parts from point cloud. ACM Trans. Graph. **38**(6), 1–15 (2019)

46. Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., Xie, S.: Representation alignment for generation: Training diffusion transformers is easier than you think. In: Int. Conf. Learn. Represent. (2025)
47. Zhu, C., Yu, R., Feng, S., Burchfiel, B., Shah, P., Gupta, A.: Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. In: Proceedings of Robotics: Science and Systems (2025)