GaussianArt: Unified Modeling of Geometry and Motion for Articulated Objects

Licheng Shen[1,2*], Saining Zhang[1,2,3*†], Honghan Li[1,2,3*], Peilin Yang[1,4],
Zihao Huang[1,5], Zongzheng Zhang[1,2], Hao Zhao[2,1‡]

[1]BAAI [2]AIR, THU [3]NTU [4]BIT [5]HUST

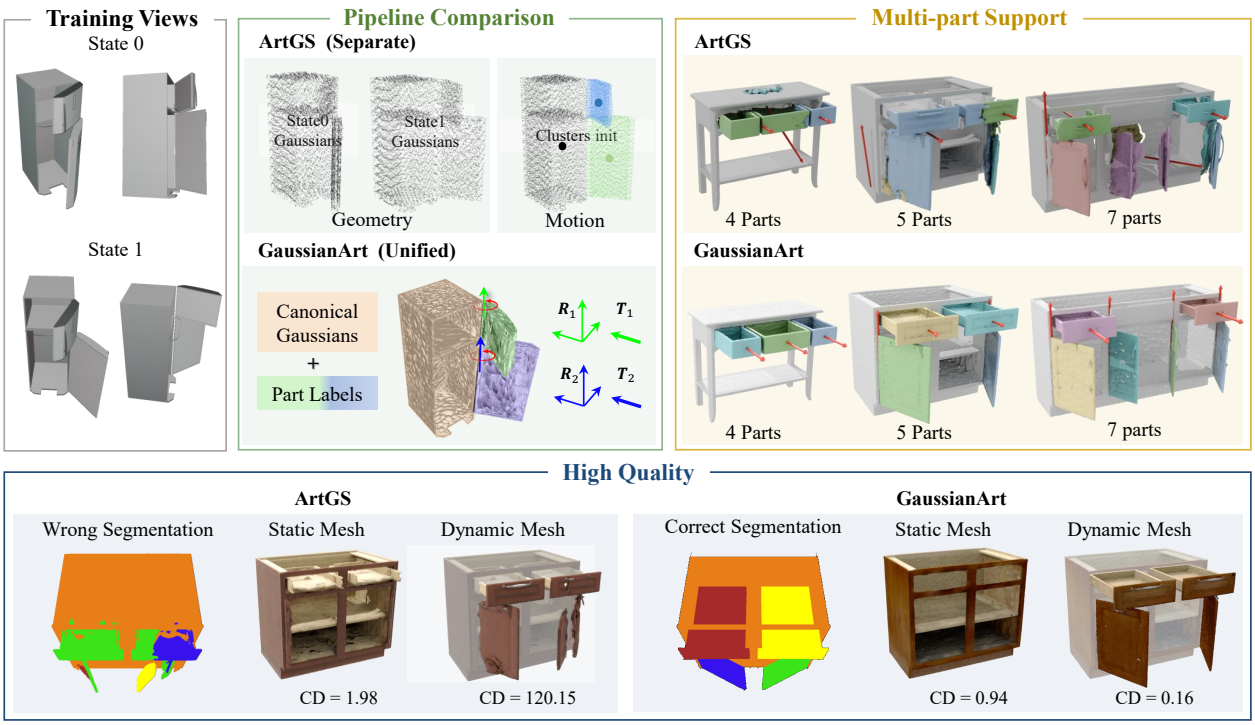

Figure 1. ArtGS adopts a separate pipeline, while GaussianArt introduces a unified representation that jointly models geometry and motion. Compared to ArtGS, which struggles with wrong part assignments and axis errors beyond 2–3 parts, our method robustly handles complex objects with many parts. In terms of quality, ArtGS produces wrong segmentation and high errors (CD = 1.98 / 120.15 for static/dynamic), while GaussianArt achieves correct segmentation and much lower errors (CD = 0.94 / 0.16). These results highlight the scalability and accuracy benefits of our unified design.

Abstract

Reconstructing articulated objects is essential for building digital twins of interactive environments. However, prior methods typically decouple geometry and motion by first reconstructing object shape in distinct states and then estimating articulation through post-hoc alignment. This separation complicates the reconstruction pipeline and restricts scalability, especially for objects with complex, multi-part articulation. We introduce a ***unified*** *representation that jointly models geometry and motion using articulated 3D Gaussians. This formulation improves robustness in motion decomposition and supports articulated objects with up to 20 parts, significantly outperforming prior approaches that often struggle beyond 2–3 parts due to brittle initialization. To systematically assess scalability and generalization, we propose MPArt-90, a new benchmark consisting of 90 articulated objects across 20 categories, each with diverse part counts and motion configurations.*

* Equal contribution.
† Project leader.
‡ Corresponding author.

Extensive experiments show that our method consistently achieves superior accuracy in part-level geometry reconstruction and motion estimation across a broad range of object types. We further demonstrate applicability to downstream tasks such as robotic simulation and human-scene interaction modeling, highlighting the potential of unified articulated representations in scalable physical modeling. Project Page.

1. Introduction

Reconstructing articulated objects plays a central role in creating digital twins for robotic simulation and generic interaction modeling [13, 36, 48]. While recent progress [12, 17, 30, 33, 38, 54, 65, 79] has been made, most existing pipelines [30, 38, 65, 79] adopt a decoupled design as shown in Fig. 1: they first reconstruct geometry from two static observations and subsequently infer motion through part-wise alignment. This separation not only introduces redundant modeling and brittle optimization, **but more critically**, breaks the physical consistency across object states—the same region of geometry is reconstructed independently per state without a coherent articulation structure.

As shown in Fig. 1 (top-left and top-middle), ArtGS [38] exemplifies this limitation: it separately models geometry and motion using per-state 3D Gaussians, then relies on clustering to estimate part motion. Such pipelines often suffer from incorrect part grouping and axis misalignment, especially when handling more than 2–3 moving parts.

Several recent systems, including DigitalTwinArt [65] and ArtGS [38], attempt to address multi-part articulation by leveraging learned part-field or clustering heuristics. However, they are constrained by: (1) separate optimization of geometry and motion; (2) brittle initialization, often failing to produce reliable part decomposition under complex configurations; and (3) limited scalability, rarely tested beyond a dozen objects, and often constrained to two-part motion. Moreover, current benchmarks are narrow in scope—most evaluate on fewer than 20 objects with constrained topology and articulation patterns.

We introduce **GaussianArt**, a physically consistent and scalable framework for articulated object reconstruction. Unlike prior methods, GaussianArt adopts a unified representation based on articulated 3D Gaussian primitives, where each Gaussian simultaneously encodes its part affiliation (via learned soft assignments) and its rigid motion (as a mixture of motion bases). This formulation enables geometry and motion to be co-optimized within a single differentiable structure, ensuring cross-state consistency and interpretability, as visualized in Fig. 1 (middle row).

To evaluate scalability, we construct **MPArt-90**, a benchmark of 90 articulated objects across 20 categories, each with observations in two states and full articulation ground truth. As shown in Fig. 1 (top-right), GaussianArt correctly recovers fine-grained part structure and motion parameters, even with 7 components. In contrast, ArtGS suffers from cluster collapse, resulting in incorrect part grouping and misaligned joints. In terms of quality, Fig. 1 (bottom row) highlights the significant gains in both segmentation accuracy and reconstruction fidelity. GaussianArt reduces the dynamic part Chamfer Distance (CD) from 120.15 to 0.16, demonstrating an order-of-magnitude improvement.

In summary, our contributions are:

- We propose GaussianArt, a reconstruction method that jointly models geometry and motion using articulated 3D Gaussians, enabling consistent reasoning across states.
- We design a soft-to-hard training paradigm that progressively refines part segmentation and rigid motion parameters, improving robustness on complex multi-part objects.
- We evaluate on MPArt-90, the largest benchmark to date for articulated object reconstruction, featuring 90 objects across 20 categories, with up to 20 parts (19 movable parts) and ground-truth motion annotations.
- Our method significantly outperforms ArtGS in both geometry and motion accuracy, and supports deployment in downstream tasks such as robotic manipulation and human-scene interaction (HSI) modeling.

2. Related Work

2.1. 3DGS and Dynamic Variants

3D Gaussian Splatting (3DGS) [21, 51, 73] represents 3D scenes using ellipsoids with Gaussian distributions as geometric primitives. This representation enables novel view rendering through differentiable rasterization, making it highly efficient for both training and rendering. The superior performance of 3DGS has inspired a series of studies that apply it to more complex dynamic scenes, exploring various motion modeling approaches [15, 32, 59, 71, 74, 75]. Several studies, such as GART [25], utilize articulated motion primitives for modeling rigid human body movement via linear blend skinning (LBS), while Shape-of-Motion [62] applies a similar model for reconstructing dynamic scenes from monocular videos. Other approaches, like SC-GS [15], use sparsely distributed control points with $\mathbf{SE}(3)$ motion priors, allowing user editing of motion, and Gaussian-Flow [32] employs a dual-domain deformation model considering frequency-domain transformations. However, most of these methods lack motion constraints and struggle to accurately estimate part-level articulated motion parameters.

2.2. 3D Articulated Object Modeling

With the advancements of shape and dynamic modeling [14, 18, 29, 40, 56, 57, 85], many works begin leveraging 3D point clouds [10, 11, 24, 28, 63, 64, 72, 77, 83, 84], or

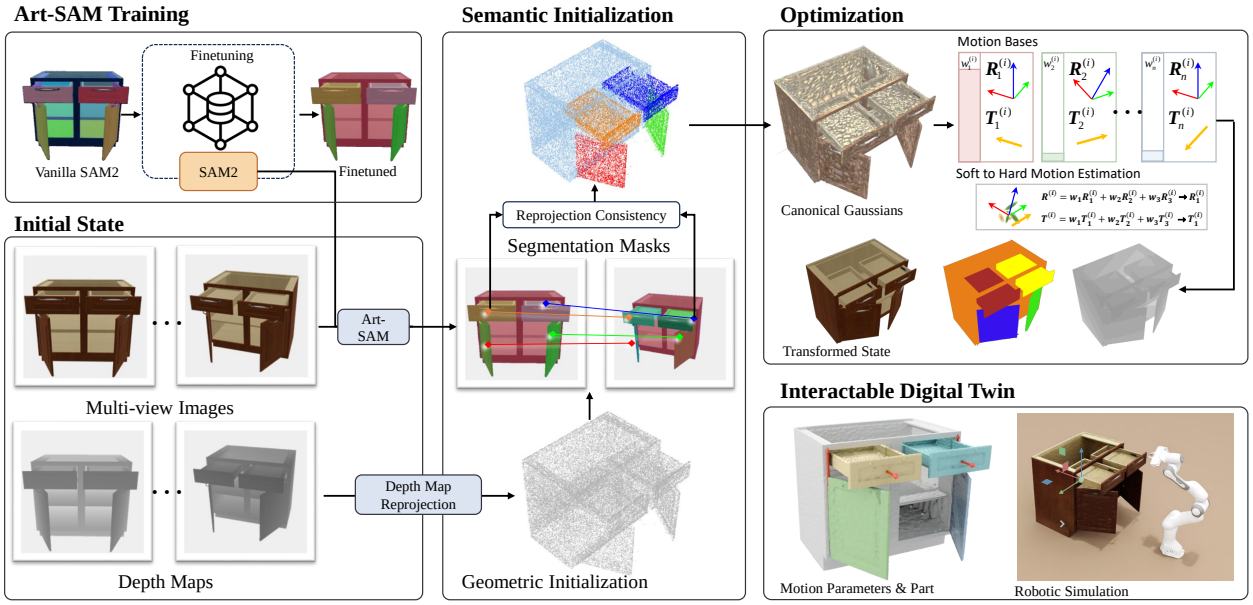

Figure 2. The overview of GaussianArt. We first design a pipeline to generate multi-view-consistent part segmentation masks, which are used to initialize Gaussians in the canonical state. During training, we introduce a unified framework that jointly learns part segmentation and motion using Gaussians. This process employs a soft-to-hard motion optimization strategy, supervised by RGB-D data and part segmentation masks, along with additional refinement techniques (see Sec. 3.3). Finally, the mesh and motion parameters produced by GaussianArt can be effectively applied to robotic simulation.

images and videos (sometimes with depth) [1, 4, 9, 10, 16, 19, 27, 35, 41, 42, 46, 49, 53, 67–69, 82] to model articulated objects. However, the most effective way is to reconstruct high-fidelity digital twins through multi-state observation [6, 12, 17, 22, 30, 33, 38, 47, 54, 55, 60, 65, 66, 79].

Former works [17, 47] use different scene representations to reconstruct articulated objects. The emergence of neural radiance fields [34, 37, 44, 61, 81] and their high rendering fidelity have led to a series of works adopting neural implicit scene representations for articulated object reconstruction. [33, 55] are among the first efforts in reconstructing articulated objects based on neural rendering. Although promising results have been achieved for two-part objects, extending these approaches to multi-part scenarios remains challenging and lacks generalizability. Recent studies [6, 54, 65] have improved the accuracy of motion parameter prediction and demonstrated some capability in multi-part object reconstruction, but they struggle to generalize to objects with more parts and more complex motion combinations. ArtGS [38], built upon 3DGS, achieves certain improvements on multi-part objects, but its results are highly unstable and show poor generalization in large-scale evaluations. In this work, we propose a well-designed unified scalable reconstruction pipeline based on articulated Gaussians.

3. GaussianArt

In this section, we present GaussianArt, a unified pipeline for modeling articulated objects using 3DGS. Sec. 3.2 provides the design and analysis of the articulated Gaussians. We employ a soft-to-hard training paradigm to optimize the Gaussians as rigid parts (Sec. 3.3) and build robust initialization (Sec. 3.5). An overview of GaussianArt is shown in Fig. 2, with key factors discussed in the following sections.

3.1. Preliminaries

3DGS [21] represents a 3D scene by a set of Gaussian Primitives, each defined as:

$$\mathbf{G}(\mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right), \quad (1)$$

where $x \in \mathbb{R}^{3\times1}$ is Gaussian's 3D position in the scene, $\boldsymbol{\mu} \in \mathbb{R}^{3\times1}$ is the mean vector, and $\boldsymbol{\Sigma} \in \mathbb{R}^{3\times3}$ is the covariance matrix. To ensure positive semi-definiteness, $\boldsymbol{\Sigma}$ is parameterized as $\boldsymbol{\Sigma} = \mathbf{RSS}^T\mathbf{R}^T$, where $\mathbf{R} \in \mathbb{R}^{3\times3}$ is a rotation matrix and $\mathbf{S} \in \mathbb{R}^{3\times3}$ is a scaling matrix.

For rendering, 3D Gaussians are depth-sorted, projected, and alpha-blended on the 2D plane to form pixel colors:

$$\mathbf{C} = \sum_{i=1}^{n} T_i\alpha_i\mathbf{c}_i, \quad T_i = \prod_{j=1}^{i-1}(1-\alpha_j), \quad (2)$$

where n is the number of contributing 2D Gaussians, T_i is the transmission factor, α_j is the opacity and $\mathbf{c}_i$ represents the spherical harmonics-based color of the i-th Gaussian. Other attributes, such as depth, normals, and even semantics, can also be rendered in this way.

To reconstruct articulated objects with multiple 1-DoF motion parts, we follow the two-state observation setting in DigitalTwinArt [65], requiring posed RGB-D sequences of the same scene in two states as input: $\left\{\overline{I}_i^t, \overline{D}_i^t, \overline{E}_i^t, \overline{K}_i^t\right\}_{i=1}^{N_s}$,

$t \in \{0, 1\}$, where N_s denotes the number of images, $\overline{I}_i^t$ the RGB image, $\overline{D}_i^t$ the depth map, $\overline{E}_i^t$ the camera extrinsics, and $\overline{K}_i^t$ the camera intrinsics..

3.2. Articulated Object Representation

While effective for static geometry, vanilla 3DGS cannot represent part-wise rigid motion or ensure cross-state consistency—each state must be modeled independently, which breaks physical plausibility in articulated systems. To address this, we extend 3DGS with an articulated formulation.

Since 3DGS is an explicit scene representation, rigid motions can be modeled by directly transforming Gaussian primitives. Thus, we reconstruct articulated objects as a motion field over canonical Gaussians. For each primitive, rigid motion applies to its mean and covariance:

$$\widetilde{\mu}^{(i)} = \mathbf{R}^{(i)}\mu^{(i)} + \mathbf{T}^{(i)}, \widetilde{\mathbf{\Sigma}}^{(i)} = \mathbf{R}^{(i)}\mathbf{\Sigma}^{(i)}\mathbf{R}^{(i)^T}, \quad (3)$$

where $\widetilde{\mu}^{(i)} \in \mathbb{R}^3, \widetilde{\mathbf{\Sigma}}^{(i)} \in \mathbb{R}^{3\times3}$ denotes the means and covariance after transformation, and $\mathbf{R}^{(i)}, \mathbf{T}^{(i)}$ denotes the rigid motion parameters of each Gaussian primitive. In articulated scenes, the number of movable parts is far fewer than primitives. We therefore define global motion bases $\{\mathbf{R}_i, \mathbf{T}_i\}_{i=1}^N$ (N stands for the number of parts), and express per-Gaussian motion as a weighted combination:

$$\mathbf{R}^{(i)} = \sum_{j=1}^{N} w_j^{(i)}\mathbf{R}_j, \mathbf{T}^{(i)} = \sum_{j=1}^{N} w_j^{(i)}\mathbf{T}_j, \quad (4)$$

where blending weights $\mathbf{w}^{(i)} = w_1^{(i)}, w_2^{(i)}, \ldots, w_N^{(i)} \in \mathbb{R}^N$ denote the probability of each primitive belonging to a motion base.

This modeling alone is insufficient for accurately capturing the characteristics of articulated objects due to the lack of constraints. To precisely model articulated objects, the following properties should be satisfied during training:

- One-hot weights: each Gaussian strongly belongs to one motion base.
- Spatial sparsity: motion weights vary only near part boundaries.
- Rigid estimation: primitives within a part are treated as a rigid body for efficient joint optimization.

3.3. Soft-to-hard Training Paradigm

This formulation encodes both geometry and motion in a differentiable manner and ensures physical consistency across object states. However, optimizing such a soft, over-parameterized motion field is challenging, especially when parts are ambiguous or occluded. To address this, GaussianArt adopts a soft-to-hard training paradigm that learns motion as rigid parts.

During the initial 6,000 iterations, we warm up the Gaussians at the canonical state. After the warm-up, since part geometry, assignments, and motion parameters are initially suboptimal, we employ a soft learning strategy to refine

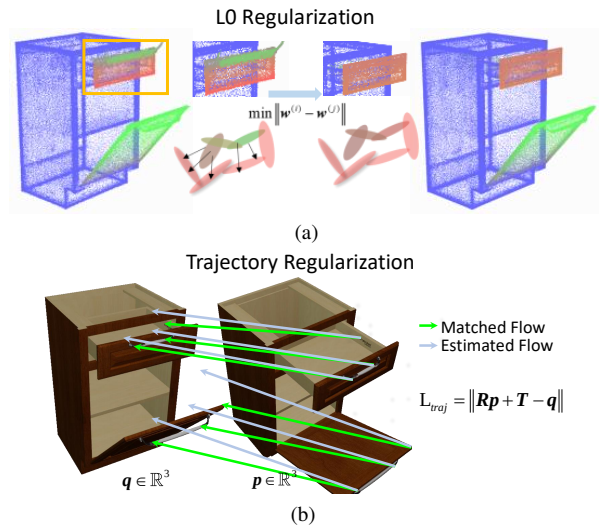

Figure 3. Regularization during training. (a) L_0 regularization: By using L_0 regularization, erroneously assigned Gaussians can be progressively corrected. (b) Trajectory regularization: The point transformed by $\mathbf{p}$ using the estimated flow is constrained to approximate the matched point, facilitating the efficient optimization of motion parameters.

Gaussian weights $\mathbf{w}^{(i)}$ and refine part geometries, which ensures rigid part segmentation while laying the foundation for hard training. Specifically, we estimate Gaussian motion as a soft mode Eq. (4) for smooth learning and use two-state part segmentation masks (Sec. 3.5) to guide $\mathbf{w}^{(i)}$ via rasterization, enforcing robust boundary constraints.

However, segmentation masks alone are insufficient for regularizing the weights of all Gaussians. To enhance regularization, we incorporate L_0 gradient sparsity during optimization:

$$\mathcal{L}_{\text{sparsity}} = \sum_{i=1}^{N} \sum_{j \in \texttt{KNN}(i)} \left\| \mathbf{w}^{(i)} - \mathbf{w}^{(j)} \right\|. \quad (5)$$

As depicted in Fig. 3a, L_0 regularization corrects misallocated Gaussians, preventing mix-up. It enforces spatial consistency, ensuring nearby Gaussians produce similar predictions and approximate rigid parts, facilitating subsequent hard training.

In other training settings, the Gaussian position learning rate decays exponentially to zero before hard training, ensuring geometric stability and focusing on motion learning. To prevent interference in prismatic parts, we classify a part as prismatic if its rotation remains below a threshold ϵ after 1,000 soft training steps. We then fix its rotation quaternion to the identity quaternion and detach it from further updates.

During hard training, we disable L_0 regularization and assign each Gaussian's motion parameters to those of the part with the highest weight:

$$\mathbf{R}^{(i)} = \mathbf{R}_{j*}, \mathbf{T}^{(i)} = \mathbf{T}_{j*}, \text{where } j^* = \arg\max_j w_j^{(i)}. \quad (6)$$

This enables simple and direct motion optimization.

Sometimes, part segmentation alone cannot fully constrain motion parameter learning, making extreme move-

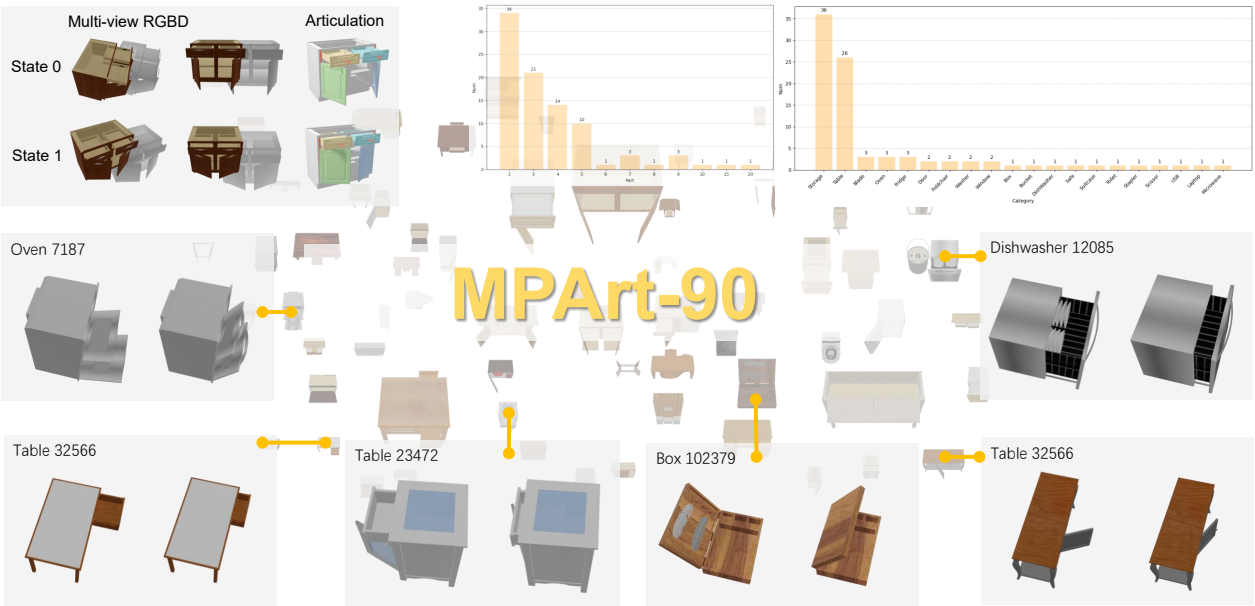

Figure 4. MPArt-90 benchmark. Unlike prior datasets that contain fewer than 20 objects and quickly saturate, MPArt-90 scales articulated object reconstruction to 90 objects across 20 diverse categories. Each object provides multi-view RGBD observations together with ground-truth motion parameters, covering configurations with up to 20 parts. This scale reveals failure cases of prior methods, which often collapse beyond 2–3 parts due to brittle initialization. By offering a large and physically grounded benchmark, MPArt-90 enables systematic evaluation of scalability and generalization in articulated modeling.

ments and occlusions difficult to model. To refine motion learning, we apply a feature-matching-based trajectory regularization term during hard training. For image I_v^0 from view v at state 0, we select the top k closest view from state 1 and apply an image feature matching model [52] to establish 2D pixel correspondences $\{(\mathbf{p}_i, \mathbf{q}_i)\}_{i=1}^{N_m}$, where N_m denotes the number of matches and $\mathbf{p}_i, \mathbf{q}_i$ are pixel coordinates. We lift them to 3D space as $\mathcal{M} = \{(\tilde{\mathbf{p}}_i, \tilde{\mathbf{q}}_i)\}_{i=1}^{N_m}$ using depth and camera parameters. Then, we apply a 3D locality filter on 3D maches to mitigate false correspondences, yielding a refined set of matching pairs $F(\mathcal{M})$. The transformation $(\mathbf{R}_j, \mathbf{T}_j)$ at the matching positions can be derived from Eq. (6). The trajectory regularization term is the average difference between the position obtained through transformation and that derived from the filtered correspondences (Fig. 3b):

$$\mathcal{L}_{traj} = \sum_{j \in F(\mathcal{M})} \left\| (\mathbf{R}_j \mathbf{p}_j + \mathbf{T}_j) - \mathbf{q}_j \right\| . \tag{7}$$

3.4. Optimization

We supervise training via a multi-term loss that balances appearance fidelity, part consistency, and physically plausible motion. $\mathcal{L}_{\text{RGB-D}}$ is the rendering loss like:

$$\mathcal{L}_{\text{RGB-D}} = (1 - \lambda_{\text{SSIM}})\mathcal{L}_1 + \lambda_{\text{SSIM}}\mathcal{L}_{\text{D-SSIM}} + \lambda_{\text{D}}\mathcal{L}_{\text{D}} \tag{8}$$

where $\mathcal{L}_1 = \left\| I - \overline{I} \right\|_1$, $\mathcal{L}_{\text{D-SSIM}}$ is the D-SSIM loss [21], and $\mathcal{L}_{\text{D}} = \left\| D - \overline{D} \right\|_1$. We also use the segmentation loss:

$$\mathcal{L}_{\text{SEM}} = \text{H}(P, \overline{P}) \tag{9}$$

where P is the rendering of weights $\mathbf{w}$, $\overline{P}$ is the segmentation mask of parts, and H is the cross-entropy loss. All in all, our supervision could be summarized as:

$$\mathcal{L}_{\text{Soft}} = \mathcal{L}_{\text{RGB-D}} + \lambda_{\text{SEM}}\mathcal{L}_{\text{SEM}} + \lambda_{\text{sparsity}}\mathcal{L}_{\text{sparsity}}, \tag{10}$$

$$\mathcal{L}_{\text{Hard}} = \mathcal{L}_{\text{RGB-D}} + \lambda_{\text{SEM}}\mathcal{L}_{\text{SEM}} + \lambda_{\text{traj}}\mathcal{L}_{\text{traj}}, \tag{11}$$

where Eq. (10) is for soft training, and Eq. (11) is for the hard. See Supplementary Material Sec. 11 for details.

3.5. Initialization

Part segmentation. To initialize and regularize Gaussian part weights $\mathbf{w}^{(i)}$, we leverage SAM2 [50], a foundation segmentation model pretrained on large-scale data. Since its zero-shot results often show inconsistent granularity on articulated objects, we fine-tune it on multi-view images and masks rendered from Partnet-Mobility (PM) [70], yielding a specialized model, Art-SAM. We then apply cross-view propagation to ensure multi-view consistency (see Supplementary Material Secs. 9 and 10). This method is robust and generalizable, as Art-SAM requires only light post-training to adapt to novel object categories or configurations.

Canonical Gaussians Initialization. First, we select the joint state with higher visibility as the canonical state. After obtaining view-consistent segmentation masks, we randomly sample points from RGB-D images and reproject both the color and the part label into 3D space using the depth map. This process yields point clouds for initializing the canonical Gaussians. The part label $S^{(i)}$ is then employed as an affinity feature $\mathbf{w}'^{(i)} = (w_1'^{(i)}, w_2'^{(i)}, \ldots, w_N'^{(i)}) \in \mathbb{R}^N$ attached to the Gaussian and serves as the initialization of the weights term $\mathbf{w}^{(i)}$ in mo-

tion estimation as follows:

$$w_j^{'(i)} = \begin{cases} 1 & \text{if } j = S^{(i)}, \\ 0 & \text{if } j \neq S^{(i)}, \end{cases} \tag{12}$$

$$\mathbf{w}^{(i)} = \text{Softmax}(\mathbf{w}^{'(i)}). \tag{13}$$

This method for part initialization is highly robust while also providing ample room for modification.

4. MPArt-90 Benchmark

4.1. Benchmark Data Generation

We have constructed a novel benchmark, MPArt-90, containing 90 objects from 20 categories, extending the existing dataset for articulated object reconstruction to a larger scale, covering more object types and motion patterns, as shown in Fig. 4. Compared to earlier benchmarks limited to fewer than 20 objects and only 2–3 part articulations, MPArt-90 emphasizes diversity and realism, exposing failure modes in brittle pipelines and enabling systematic evaluation of generalization and robustness.

Articulated objects are mainly constructed from the PM dataset [70], from which we select 87 objects based on the diversity of categories, part number, and appearances. We use Blender with a procedural rendering pipeline [5, 7] to render multi-view images of the 3D models in the base assets. To increase the diversity of object states while providing adequate observation of the interior parts of the objects, we set the 1-DoF part-level motion parameter to two random states: the starting state lies between $[0.65, 0.75]$ and the ending state is within $[0.35, 0.45]$ (here 0 denotes the "fully-closed" state and 1 denotes the "fully-open" state). For the object at each motion state, we place the camera in a spherical region around the object and randomly sample 100 views for training and 20 views for testing, all at a resolution of 800×800.

Due to the limited availability of high-quality real-world articulated object data and the significant ground-truth errors caused by the inability to annotate the internal structures of real objects, we selected only three well-annotated real objects from the Multiscan [43] dataset. See Supplementary Material Sec. 18 for more results.

5. Experimental Results

5.1. Implementations

We first evaluate several methods on several traditional datasets, and finally select our method and ArtGS [38] for scale-up evaluation on our MPArt-90 benchmark.

For metrics, we calculate Axis Pos Error, Axis Angle Error, and Part Motion Error for motion parameters estimation, and CD for static and dynamic part separately for geometric reconstruction. We report the mean for each metric over the 3 trials at the high-visibility state. See Supplementary Material Secs. 15 and 16 for more details.

5.2. Experiments on MPArt-90

As depicted in Tab. 1, when the object contains fewer parts, ArtGS performs comparably to, or even slightly better than, GaussianArt. However, as the number of parts increases, the stability of ArtGS degrades and the performance gap between the two methods widens. This is particularly evident in the geometry of dynamic parts, where ArtGS's cluster-based initialization lacks strong generalizability. In many cases, it fails to achieve effective part segmentation, resulting in significant errors in motion parameter estimation. In contrast, GaussianArt uses a fine-tuned vision foundation model and a multi-view consistency pipeline to produce high-quality part segmentation, providing stable initialization for motion parameter learning and improving generalization to diverse objects.

Moreover, as illustrated by the qualitative results in Fig. 5, when handling objects with multiple parts exhibiting similar motion patterns, ArtGS often produces ambiguous segmentation results, which in turn lead to inaccurate motion parameter estimation. Such errors frequently cause parts to split during motion, adversely affecting the overall reconstructed geometry. In more complex scenarios, such as objects comprising more than 20 parts, ArtGS may even fail to converge during training. In contrast, GaussianArt exhibits strong generalization capabilities: even when confronted with highly articulated objects, it consistently delivers accurate part segmentation and reliable motion parameter predictions. These strengths enable the reconstruction of high-fidelity digital twins, laying a solid foundation for wider deployment in downstream applications. See Supplementary Material Sec. 17 for more results.

5.3. Ablation Studies

Implementation Details. To evaluate the effectiveness of different components, we design several ablation studies on 5 multi-part objects. All metrics are averaged across 5 trials.

Results. The results are as follows:

- *Part assignment.* We evaluate our model under 3 different scenarios: (1) without part segmentation masks, (2) without part initialization, and (3) replacing the part assignment module with an MLP, as used in GART [25]. As shown in Table 2, without segmentation masks, Gaussians struggle to capture the complex motion of multiple parts. Additionally, relying solely on part segmentation supervision—without explicit part initialization—can negatively impact the geometric reconstruction of movable parts. When using MLPs for part assignment, the model fails entirely to learn part segmentation and motion for multi-part objects.
- *Trajectory regularization.* As demonstrated in Table 2, trajectory regularization enhances part motion learning, improving articulated reconstruction.

		2 Parts (34)	3 Parts (21)	4-5 Parts (24)	6-20 Parts (11)	All (90)
Axis Ang	ArtGS [38]	**3.53**	11.61	15.49	35.66	24.34
	Ours	4.90	**6.33**	**12.43**	**12.05**	**12.17**
Axis Pos	ArtGS [38]	1.08	1.62	**1.09**	4.62	1.45
	Ours	**0.27**	**0.38**	1.30	**3.06**	**1.06**
Part Motion	ArtGS [38]	7.98	**6.08**	**4.45**	13.74	10.16
	Ours	**7.57**	11.82	9.03	**7.14**	**9.07**
CD-s	ArtGS [38]	5.39	13.37	18.02	13.14	11.57
	Ours	**2.75**	**2.47**	**2.31**	**3.70**	**2.68**
CD-m	ArtGS [38]	47.80	194.15	340.53	459.81	380.29
	Ours	**4.61**	**6.17**	**5.42**	**5.43**	**5.46**

Table 1. Quantitative results on MPArt-90 benchmark. Metrics are shown as the mean $\pm$ std over 3 trials with different random seeds following [65].

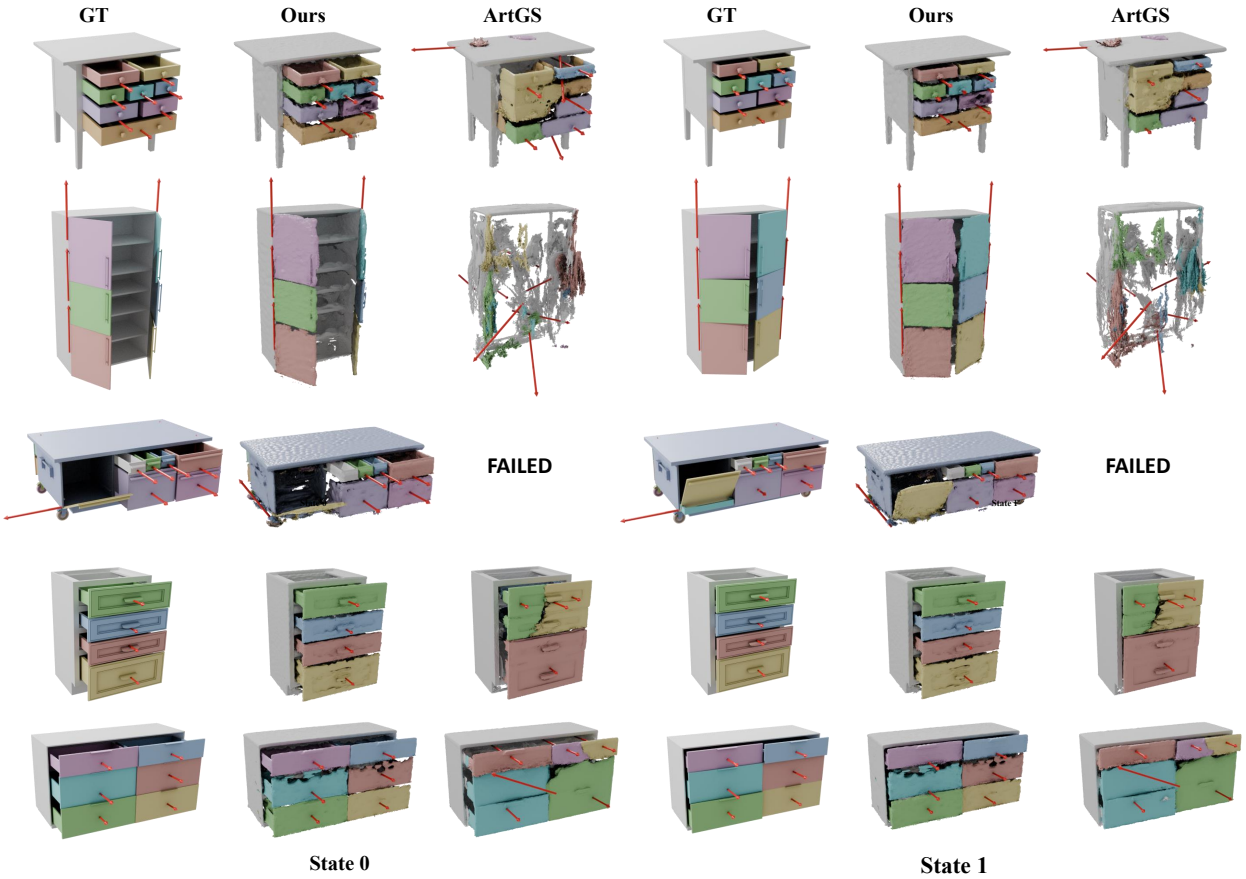

Figure 5. Qualitative results on multi-part objects of MPArt-90.

- *L_0 regularization.* As observed in Table 2, L_0 regularization refines part assignment, leading to more accurate part articulation modeling.

6. Application

6.1. Robotic Manipulation

Fig. 6 presents GaussianArt's reconstruction of multi-part articulated objects in NVIDIA Omniverse IssacSim. Leveraging learned motion parameters and precise part-level geometry, we effectively decomposed the hybrid motion in vi-

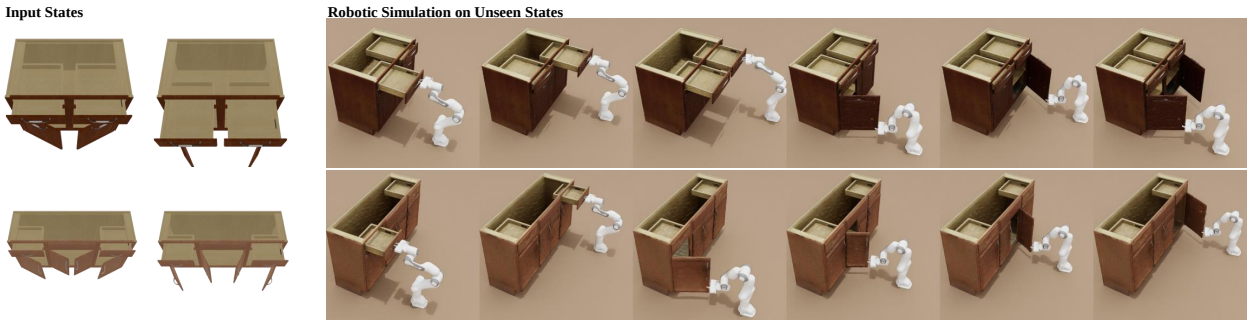

Figure 6. Digital twins reconstructed by GaussianArt in NVIDIA Omniverse IssacSim.

	Axis Ang	Axis Pos	Part Motion	CD-s	CD-m
Proposed	**0.03**	**0.01**	**0.04**	**0.67**	**0.14**
w/o Part-seg	28.60	4.67	18.50	0.88	186.67
w/o L_0	0.10	0.01	0.08	0.70	0.17
w/o Traj	0.31	0.03	0.35	0.85	0.25
w/o Part-init	0.25	0.02	0.15	0.71	1.34
w MLP Seg	41.57	3.78	31.43	2.68	478.20

Table 2. Results of ablation studies.

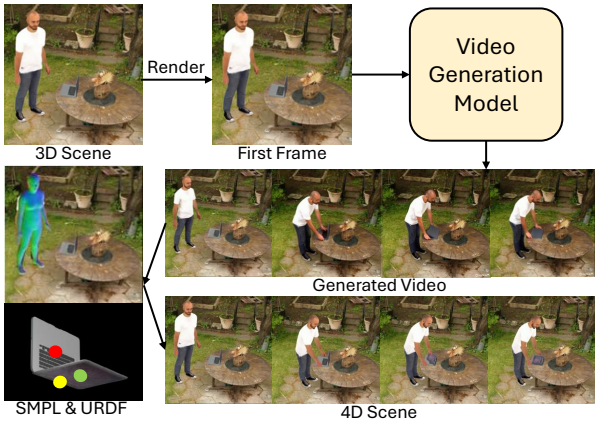

Figure 7. HSI pipeline with the digital twin generated by GaussianArt.

sual observations, enabling a robot arm to interact with any moving part at unseen states in input images. These realistic digital twins facilitate robotic manipulation of articulated objects.

6.2. HSI

Furthermore, our Articulated Gaussians can be leveraged to generate 4D assets, enabling the modeling of HSI in dynamic environments. As depicted in Fig. 7, given Gaussian-based representations of humans, objects, and scenes as input, inspired by ZeroHSI [26], we could generate high-fidelity 4D scenes. We first render a static frame, which is then used as input to a text-guided video diffusion model to synthesize a corresponding video. This generated video serves as a supervisory signal to optimize the kinematic parameters of humans and objects, effectively lifting the original 3D assets into a coherent 4D representation.

Specifically, we employ SMPL [39] for human kinematic modeling and use our reconstructed digital twins, with articulation parameters, to represent objects with kinematics. Through a distillation process, the motion dynamics captured in the video are transferred to the 4D scene.

We believe this approach holds strong potential for a wide range of applications and will offer significant value across multiple domains in the future.

7. Limitations

Despite its strengths, GaussianArt has several limitations. First, the lack of direct constraints on intermediate motion states can lead to incorrect motion parameter learning, particularly in extreme transitions (e.g., from a fully open to a fully closed door). Future work will explore constraint-based strategies within the unified GS framework to improve motion learning. Second, the initialization of canonical Gaussians may be suboptimal due to out-of-distribution issues in part segmentation or misalignments in multi-view reconstruction. These imperfections can negatively impact the learning of motion parameters. Addressing this challenge will require developing more robust multi-view segmentation methods. See Supplementary Material Sec. 14 for more details.

8. Conclusion

In this work, we introduce GaussianArt, a unified modeling pipeline for articulated objects. Our approach begins with a robust part segmentation model to initialize canonical Gaussians, followed by a soft-to-hard training paradigm for improved motion optimization. Extensive experiments show that GaussianArt achieves state-of-the-art (SoTA) performance in geometric reconstruction and part motion estimation on our curated largest benchmark for articulated objects reconstruction, MPArt-90. The digital twins created by GaussianArt can be seamlessly integrated into simulators for tasks like articulated object manipulation, and can also be used for modeling human–scene interactions, where we hope to enable more natural and adaptive simulation of real-world environments.

References

[1] Ziang Cao, Zhaoxi Chen, Linag Pan, and Ziwei Liu. Physx: Physical-grounded 3d asset generation. *arXiv preprint arXiv:2507.12465*, 2025. 3

[2] Jiazhong Cen, Zanwei Zhou, Jiemin Fang, chen yang, Wei Shen, Lingxi Xie, Dongsheng Jiang, XIAOPENG ZHANG, and Qi Tian. Segment anything in 3d with nerfs. In *Advances in Neural Information Processing Systems*, pages 25971–25990. Curran Associates, Inc., 2023. 2

[3] Jiazhong Cen, Jiemin Fang, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment any 3d gaussians, 2024. 1

[4] 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. 3

[5] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 6, 1

[6] Jianning Deng, Kartic Subr, and Hakan Bilen. Articulate your neRF: Unsupervised articulated object modeling via conditional view synthesis. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3

[7] Maximilian Denninger, Dominik Winkelbauer, Martin Sundermeyer, Wout Boerdijk, Markus Knauer, Klaus H. Strobl, Matthias Humt, and Rudolph Triebel. Blenderproc2: A procedural pipeline for photorealistic rendering. *Journal of Open Source Software*, 8(82):4901, 2023. 6, 1

[8] M ESTER. A density-based algorithm for discovering clusters in sarge spatial databases with noise. In *Pro. of 2nd Int. Conf. Knowledge Discovery and Data Mining, 1996*, pages 291–316, 1996. 1

[9] Mingju Gao, Yike Pan, Huan-ang Gao, Zongzheng Zhang, Wenyi Li, Hao Dong, Hao Tang, Li Yi, and Hao Zhao. Partrm: Modeling part-level dynamics with large cross-state reconstruction model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7004–7014, 2025. 3

[10] Haoran Geng, Songlin Wei, Congyue Deng, Bokui Shen, He Wang, and Leonidas Guibas. Sage: Bridging semantic and actionable parts for generalizable manipulation of articulated objects. *arXiv preprint arXiv:2312.01307*, 2023. 2, 3

[11] 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. 2

[12] Junfu Guo, Yu Xin, Gaoyi Liu, Kai Xu, Ligang Liu, and Ruizhen Hu. Articulatedgs: Self-supervised digital twin modeling of articulated objects using 3d gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27144–27153, 2025. 2, 3

[13] Ruizhen Hu, Manolis Savva, and Oliver van Kaick. Functionality representations and applications for shape analysis. In *Computer Graphics Forum*, pages 603–624. Wiley Online Library, 2018. 2

[14] Qixing Huang, Vladlen Koltun, and Leonidas Guibas. Joint shape segmentation with linear programming. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, pages 1–12, 2011. 2

[15] Yi-Hua Huang, Yang-Tian Sun, Ziyi Yang, Xiaoyang Lyu, Yan-Pei Cao, and Xiaojuan Qi. Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4220–4230, 2024. 2

[16] Hanxiao Jiang, Yongsen Mao, Manolis Savva, and Angel X. Chang. Opd: Single-view 3d openable part detection. In *Computer Vision – ECCV 2022*, pages 410–426, Cham, 2022. Springer Nature Switzerland. 3

[17] Zhenyu Jiang, Cheng-Chun Hsu, and Yuke Zhu. Ditto: Building digital twins of articulated objects from interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5616–5626, 2022. 2, 3, 4, 5

[18] R Kenny Jones, Theresa Barton, Xianghao Xu, Kai Wang, Ellen Jiang, Paul Guerrero, Niloy J Mitra, and Daniel Ritchie. Shapeassembly: Learning to generate programs for 3d shape structure synthesis. *ACM Transactions on Graphics (TOG)*, 39(6):1–20, 2020. 2

[19] Yuki Kawana and Tatsuya Harada. Detection based part-level articulated object reconstruction from single rgbd image. In *Advances in Neural Information Processing Systems*, pages 18444–18473. Curran Associates, Inc., 2023. 3

[20] Nikhil Keetha, Jay Karhade, Krishna Murthy Jatavallabhula, Gengshan Yang, Sebastian Scherer, Deva Ramanan, and Jonathon Luiten. Splatam: Splat track & map 3d gaussians for dense rgb-d slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21357–21366, 2024. 3

[21] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42(4):139:1 – 139:14, 2023. 2, 3, 5

[22] Seungyeon Kim, Junsu Ha, Young Hun Kim, Yonghyeon Lee, and Frank C Park. Screwsplat: An end-to-end method for articulated object recognition. *arXiv preprint arXiv:2508.02146*, 2025. 3

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1

[24] Jens U Kreber and Joerg Stueckler. Guiding diffusion-based articulated object generation by partial point cloud alignment and physical plausibility constraints. *arXiv preprint arXiv:2508.00558*, 2025. 2

[25] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19876–19887, 2024. 2, 6

[26] Hongjie Li, Hong-Xing Yu, Jiaman Li, and Jiajun Wu. Zerohsi: Zero-shot 4d human-scene interaction by video generation. *arXiv preprint arXiv:2412.18600*, 2024. 8

[27] Siqi Li, Xiaoxue Chen, Haoyu Cheng, Guyue Zhou, Hao Zhao, and Guanzhong Tian. Locate n' rotate: Two-stage openable part detection with foundation model priors. In *Computer Vision – ACCV 2024*, pages 93–108, Singapore, 2025. Springer Nature Singapore. 3

[28] Xiaolong Li, He Wang, Li Yi, Leonidas J. Guibas, A. Lynn Abbott, and Shuran Song. Category-level articulated object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[29] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10486–10496, 2025. 2

[30] Shengjie Lin, Jiading Fang, Muhammad Zubair Irshad, Vitor Campagnolo Guizilini, Rares Andrei Ambrus, Greg Shakhnarovich, and Matthew R Walter. Splart: Articulation estimation and part-level reconstruction with 3d gaussian splatting. *arXiv preprint arXiv:2506.03594*, 2025. 2, 3

[31] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[32] Youtian Lin, Zuozhuo Dai, Siyu Zhu, and Yao Yao. Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21136–21145, 2024. 2

[33] 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 (ICCV)*, pages 352–363, 2023. 2, 3, 4, 5, 6, 7

[34] Junchen Liu, Wenbo Hu, Zhuo Yang, Jianteng Chen, Guoliang Wang, Xiaoxue Chen, Yantong Cai, Huan-ang Gao, and Hao Zhao. Rip-nerf: Anti-aliasing radiance fields with ripmap-encoded platonic solids. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3

[35] 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. 3

[36] Jiayi Liu, Manolis Savva, and Ali Mahdavi-Amiri. Survey on modeling of articulated objects. *arXiv preprint arXiv:2403.14937*, 2024. 2

[37] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33:15651–15663, 2020. 3

[38] Yu Liu, Baoxiong Jia, Ruijie Lu, Junfeng Ni, Song-Chun Zhu, and Siyuan Huang. Artgs: Building interactable replicas of complex articulated objects via gaussian splatting. *arXiv preprint arXiv:2502.19459*, 2025. 2, 3, 6, 7, 4, 5

[39] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 8

[40] Jiaxin Lu, Yifan Sun, and Qixing Huang. Jigsaw: Learning to assemble multiple fractured objects. *Advances in Neural Information Processing Systems*, 36:14969–14986, 2023. 2

[41] Ruijie Lu, Yu Liu, Jiaxiang Tang, Junfeng Ni, Yuxiang Wang, Diwen Wan, Gang Zeng, Yixin Chen, and Siyuan Huang. Dreamart: Generating interactable articulated objects from a single image. *arXiv preprint arXiv:2507.05763*, 2025. 3

[42] Zhao Mandi, Yijia Weng, Dominik Bauer, and Shuran Song. Real2code: Reconstruct articulated objects via code generation. *arXiv preprint arXiv:2406.08474*, 2024. 3

[43] Yongsen Mao, Yiming Zhang, Hanxiao Jiang, Angel X Chang, and Manolis Savva. Multiscan: Scalable rgbd scanning for 3d environments with articulated objects. In *Advances in Neural Information Processing Systems*, 2022. 6, 1

[44] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Computer Vision – ECCV 2020*, pages 405–421, Cham, 2020. Springer International Publishing. 3

[45] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 1

[46] 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. 3

[47] Jiteng Mu, Weichao Qiu, Adam Kortylewski, Alan Yuille, Nuno Vasconcelos, and Xiaolong Wang. A-sdf: Learning disentangled signed distance functions for articulated shape representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13001–13011, 2021. 3

[48] Petra Pejić, Valentin Šimundić, Matej Džijan, and Robert Cupec. Articulated objects: From detection to manipulation—survey. In *International Conference on Intelligent Autonomous Systems*, pages 495–508. Springer, 2022. 2

[49] Shengyi Qian, Linyi Jin, Chris Rockwell, Siyi Chen, and David F. Fouhey. Understanding 3d object articulation in internet videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1599–1609, 2022. 3

[50] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feicht-

enhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 5, 1

[51] Xiaowei Song, Jv Zheng, Shiran Yuan, Huan-ang Gao, Jingwei Zhao, Xiang He, Weihao Gu, and Hao Zhao. Sags: Scale-adaptive gaussian splatting for training-free anti-aliasing. *arXiv preprint arXiv:2403.19615*, 2024. 2

[52] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 5, 3

[53] Xiaohao Sun, Hanxiao Jiang, Manolis Savva, and Angel Chang. OPDMulti: Openable Part Detection for Multiple Objects . In *2024 International Conference on 3D Vision (3DV)*, pages 169–178, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 3

[54] Archana Swaminathan, Anubhav Gupta, Kamal Gupta, Shishira R. Maiya, Vatsal Agarwal, and Abhinav Shrivastava. Leia: Latent view-invariant embeddings for implicit 3d articulation. In *Computer Vision – ECCV 2024*, pages 210–227, Cham, 2025. Springer Nature Switzerland. 2, 3

[55] 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, 2022. 3

[56] Shubham Tulsiani, Abhishek Kar, Joao Carreira, and Jitendra Malik. Learning category-specific deformable 3d models for object reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):719–731, 2016. 2

[57] Shubham Tulsiani, Hao Su, Leonidas J Guibas, Alexei A Efros, and Jitendra Malik. Learning shape abstractions by assembling volumetric primitives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2635–2643, 2017. 2

[58] Ignacio Vizzo, Tiziano Guadagnino, Jens Behley, and Cyrill Stachniss. Vdbfusion: Flexible and efficient tsdf integration of range sensor data. *Sensors*, 22(3):1296, 2022. 3

[59] Diwen Wan, Yuxiang Wang, Ruijie Lu, and Gang Zeng. Template-free articulated gaussian splatting for real-time reposable dynamic view synthesis. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2

[60] Haowen Wang, Xiaoping Yuan, Zhao Jin, Zhen Zhao, Zhengping Che, Yousong Xue, Jin Tian, Yakun Huang, and Jian Tang. Self-supervised multi-part articulated objects modeling via deformable gaussian splatting and progressive primitive segmentation. *arXiv preprint arXiv:2506.09663*, 2025. 3

[61] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 3

[62] Qianqian Wang, Vickie Ye, Hang Gao, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. 2024. 2

[63] 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. 2

[64] Yijia Weng, He Wang, Qiang Zhou, Yuzhe Qin, Yueqi Duan, Qingnan Fan, Baoquan Chen, Hao Su, and Leonidas J. Guibas. Captra: Category-level pose tracking for rigid and articulated objects from point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13209–13218, 2021. 2

[65] 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 (CVPR)*, pages 3141–3150, 2024. 2, 3, 7, 4, 5, 6

[66] Di Wu, Liu Liu, Zhou Linli, Anran Huang, Liangtu Song, Qiaojun Yu, Qi Wu, and Cewu Lu. Reartgs: Reconstructing and generating articulated objects via 3d gaussian splatting with geometric and motion constraints. *arXiv preprint arXiv:2503.06677*, 2025. 3

[67] Mingxuan Wu, Huang Huang, Justin Kerr, Chung Min Kim, Anthony Zhang, Brent Yi, and Angjoo Kanazawa. Predict-optimize-distill: A self-improving cycle for 4d object understanding. *arXiv preprint arXiv:2504.17441*, 2025. 3

[68] Ruiqi Wu, Xinjie Wang, Liu Liu, Chunle Guo, Jiaxiong Qiu, Chongyi Li, Lichao Huang, Zhizhong Su, and Ming-Ming Cheng. Dipo: Dual-state images controlled articulated object generation powered by diverse data. *arXiv preprint arXiv:2505.20460*, 2025.

[69] Hongchi Xia, Entong Su, Marius Memmel, Arhan Jain, Raymond Yu, Numfor Mbiziwo-Tiapo, Ali Farhadi, Abhishek Gupta, Shenlong Wang, and Wei-Chiu Ma. Drawer: Digital reconstruction and articulation with environment realism. *arXiv preprint arXiv:2504.15278*, 2025. 3

[70] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. Sapien: A simulated part-based interactive environment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5, 6, 1

[71] Zhen Xu, Yinghao Xu, Zhiyuan Yu, Sida Peng, Jiaming Sun, Hujun Bao, and Xiaowei Zhou. Representing long volumetric video with temporal gaussian hierarchy. *ACM Transactions on Graphics*, 43(6), 2024. 2

[72] 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. 2

[73] Runyi Yang, Zhenxin Zhu, Zhou Jiang, Baijun Ye, Xiaoxue Chen, Yifei Zhang, Yuantao Chen, Jian Zhao, and Hao Zhao. Spectrally pruned gaussian fields with neural compensation. *arXiv preprint arXiv:2405.00676*, 2024. 2

[74] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20331–20341, 2024. 2

[75] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. In *International Conference on Learning Representations (ICLR)*, 2024. 2

[76] Chongjie Ye, Yinyu Nie, Jiahao Chang, Yuantao Chen, Yihao Zhi, and Xiaoguang Han. Gaustudio: A modular framework for 3d gaussian splatting and beyond. *arXiv preprint arXiv:2403.19632*, 2024. 3

[77] Li Yi, Haibin Huang, Difan Liu, Evangelos Kalogerakis, Hao Su, and Leonidas Guibas. Deep part induction from articulated object pairs. *ACM Trans. Graph.*, 37(6), 2018. 2

[78] Haiyang Ying, Yixuan Yin, Jinzhi Zhang, Fan Wang, Tao Yu, Ruqi Huang, and Lu Fang. Omniseg3d: Omniversal 3d segmentation via hierarchical contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20612–20622, 2024. 1

[79] Tianjiao Yu, Vedant Shah, Muntasir Wahed, Ying Shen, Kiet A Nguyen, and Ismini Lourentzou. Part^{2}gs: Part-aware modeling of articulated objects using 3d gaussian splatting. *arXiv preprint arXiv:2506.17212*, 2025. 2, 3

[80] Zehao Yu, Torsten Sattler, and Andreas Geiger. Gaussian opacity fields: Efficient and compact surface reconstruction in unbounded scenes. *arXiv preprint arXiv:2404.10772*, 2024. 4

[81] Shiran Yuan and Hao Zhao. Slimmerf: Slimmable radiance fields. In *2024 International Conference on 3D Vision (3DV)*, pages 64–74. IEEE, 2024. 3

[82] Can Zhang and Gim Hee Lee. Iaao: Interactive affordance learning for articulated objects in 3d environments. *arXiv preprint arXiv:2504.06827*, 2025. 3

[83] Chengliang Zhong, Peixing You, Xiaoxue Chen, Hao Zhao, Fuchun Sun, Guyue Zhou, Xiaodong Mu, Chuang Gan, and Wenbing Huang. Snake: Shape-aware neural 3d keypoint field. *Advances in Neural Information Processing Systems*, 35:7052–7064, 2022. 2

[84] Chengliang Zhong, Yuhang Zheng, Yupeng Zheng, Hao Zhao, Li Yi, Xiaodong Mu, Ling Wang, Pengfei Li, Guyue Zhou, Chao Yang, et al. 3d implicit transporter for temporally consistent keypoint discovery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3869–3880, 2023. 2

[85] Chuhang Zou, Ersin Yumer, Jimei Yang, Duygu Ceylan, and Derek Hoiem. 3d-prnn: Generating shape primitives with recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 900–909, 2017. 2