BimArt: A Unified Approach for the Synthesis of 3D Bimanual Interaction with Articulated Objects

Wanyue Zhang[1,2]   Rishabh Dabral[1,2]   Vladislav Golyanik[1]   Vasileios Choutas[3]
Eduardo Alvarado[1]   Thabo Beeler[3]   Marc Habermann[1,2]   Christian Theobalt[1,2]

[1]Max Planck Institute for Informatics, Saarland Informatics Campus
[2]Saarbrücken Research Center for Visual Computing, Interaction and AI
[3]Google, Switzerland
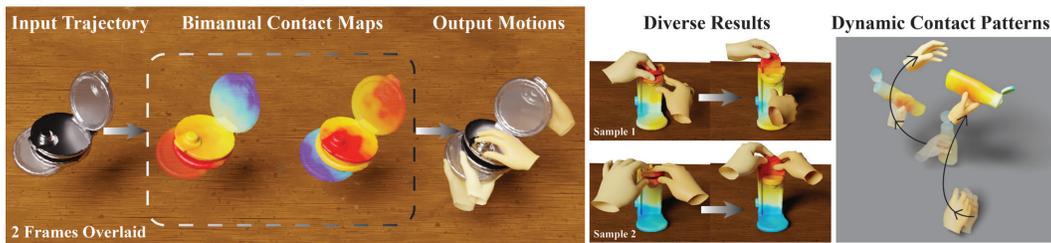
Figure 1. Given the mesh of an articulated object and its 7 DoF trajectories with 6D global states and 1D articulation, BimArt generates diverse and plausible hand motions that justify the object's trajectory. Distance-based contact maps act as intermediate features for hand-object interaction, enabling our method to generate diverse and realistic bimanual motions.

Abstract

We present BimArt, a novel generative approach for synthesizing 3D bimanual hand interactions with articulated objects. Unlike prior works, we do not rely on a reference grasp, a coarse hand trajectory, or separate modes for grasping and articulating. To achieve this, we first generate distance-based contact maps conditioned on the object trajectory with an articulation-aware feature representation, revealing rich bimanual patterns for manipulation. The learned contact prior is then used to guide our hand motion generator, producing diverse and realistic bimanual motions for object movement and articulation. Our work offers key insights into feature representation and contact prior for articulated objects, demonstrating their effectiveness in taming the complex, high-dimensional space of bimanual hand-object interactions. Through comprehensive quantitative experiments, we demonstrate a clear step towards simplified and high-quality hand-object animations that excel over the state-of-the-art in motion quality and diversity.

1. Introduction

Humans engage with articulated objects in countless ways throughout the day, whether it is twisting the cap of a water bottle, tilting a laptop screen for better viewing, or deftly slicing through paper with a pair of scissors. Although these interactions seem effortless for humans, they are challenging to generate computationally due to the highly complex and high-dimensional space of bimanual hand animations that not only rigidly move an object, but also generate meaningful object articulations. From a 3D modeling perspective, these motions require a deeper understanding of the individual object parts, their interaction affordances, and their geometry.

Despite substantial progress in 3D character animation research [20, 35, 56, 81] driven by deep learning and generative models, recent works either focus on synthesizing whole-body without considering hands [56, 69], or generate hand-object interaction assuming objects are *rigid* [9, 16, 77]. Very few studies address 3D bimanual interactions with *articulated* objects. Methods that are designed for articulated objects either work in a category-specific manner [82] for unimanual motions, cannot simultaneously perform articulation and object root translation and rotation [78], or rely on noisy hand-object interaction sequences as input and refine hand motions afterwards [39]. Some works [16, 78, 82] also assume the initial or goal grasp to be known, which can be a restrictive assumption for non-expert users.

In contrast to prior works, BimArt operates with relaxed assumptions: it does not assume a known reference grasp, is

1

not trained in an object-specific manner, does not require a coarse hand trajectory, and can perform object articulation simultaneously with the object's root rotation and translation (see a conceptual comparison in Tab. 4 in the supplementary material). Given object trajectories, which involve global translation, rotation, and articulation, **BimArt** generates diverse and realistic bimanual motions for grasping and articulating the object (see Fig. 1). We propose a three-stage approach: Our Bimanual Contact Generation model first generates contact maps, capturing dynamic interactions between the hand and the object over time. Next, the generated contact maps and the object geometry are used as conditionings to synthesize hand animations using our generative Bimanual Motion Model. Finally, we refine the generated animations with contact guidance, followed by explicit optimization to remove artifacts like penetration or missing hand-object contact.

To be category agnostic and to accommodate a wide array of geometries, we propose a novel articulation-aware representation based on basis point sets (BPS) [50], originally defined as a collection of vectors from a fixed set of points in space to the nearest vertices of the object. Our representation involves normalizing the object's scale and then computing the distance vectors from the BPS to each articulated part independently. This part-based representation treats each component of an object equally, ensuring that different surface areas have similar spatial encoding resolution. Given the above object encoding, our contact generation network predicts distance-based bimanual contact maps, which serve as an intermediate generation target, removing the need for a reference grasp. Our key insight is that frame-wise contact maps embedded on the object capture diverse grasping patterns and offer more nuanced and detailed information compared to sparser or stage-wise contact points [35, 47, 82] for bimanual interaction synthesis. We evaluate BimArt on ARCTIC [14] and HOI4D [40] datasets and achieve state-of-the-art performance in terms of interaction plausibility and diversity.

To summarize, our contributions are as follows:
- BimArt, a new approach for bimanual hand motion synthesis for interaction with articulated objects;
- A canonicalized and part-aware object feature representation, which is able to encode diverse and articulated objects in a unified representation well suited for object-aware hand animation synthesis;
- A generative model for bimanual contact maps that serve as an interaction prior for our hand motion synthesizer.

2. Related Work

3D Human Motion Synthesis. 3D human motion synthesis is an active and long-standing research field [1, 3, 4, 15, 25, 34, 44, 48, 49, 57, 67, 80]. Over the last years, neural-network-based methods have dominated it, aided by the availability of large-scale datasets of body-only [28, 43], hands-only [6, 18, 45], or whole-body motion [14, 58]. Diffusion models have manifested their potential to generate diverse and high-quality motions using conditioning signals such as text, audio, scene context, or the movements of other people [11, 17, 31, 35, 36, 46, 49, 61, 69, 73]. All these methods synthesize full-body, while hand-object interaction requires more fine-grained consideration of joint movement and alignment with object geometry.

Hand-Object Interaction. Similar to 3D human motion synthesis, the introduction of hand-object interaction datasets [2, 14, 40, 41, 58] has led to rapid developments in 3D hand and object pose reconstruction [13, 70, 71, 75, 85], static grasp synthesis [29, 30, 32, 38, 60, 62, 72], hand-object interaction (HOI) motion denoising [19, 39, 83], and dexterous object manipulation in robotics [8, 27, 33, 63, 65, 66, 68, 74]. However, existing methods [5, 16, 53, 59, 76, 82, 84] either generate single hand motions, do not work with articulated objects [16, 53, 59, 76, 84], or rely on different input assumptions such as hand trajectories [76] or textual task descriptions [5, 10, 47]. Among works that show applicability in articulated objects, text conditioning [5] lacks fine-grained control over object paths that is often essential in artistic creation. ArtiGrasp [78] requires a reference pose and cannot handle grasping and articulation simultaneously. CAMS [82] relies on the initial grasp as input and trains a separate model per category. In contrast, we train a unified model for all categories and do not rely on reference poses.

HOI Feature Representation. Existing HOI feature representations [5, 72, 76, 82, 84] are either not suitable for motion synthesis or fail to emphasize the articulated structure of objects, which is our focus. ManipNet [76] utilizes a coarse voxel-based representation to capture the object's global geometry for rigid objects. CAMS [82]'s stage-wise contact target design struggles to capture the rich bimanual interaction patterns. Works focusing on motion denoising [39, 83] compute detailed spatiotemporal features, such as motion velocities and contact correspondence. However, generating these features from scratch without assuming an initial motion is challenging and may overconstrain the synthesis model, leading to lower diversity. In contrast to these previous works, we propose a part-based object representation specifically designed for articulated objects, ensuring that objects with unbalanced part sizes are not disadvantaged. Additionally, our hand representation encodes both surface positions and distances to the object, enhancing interaction plausibility.

3. Method

Our goal is to generate realistic, diverse, and contact-aware 3D bimanual motion from a sequence of articulated object states. We consider two-part articulated objects with a total
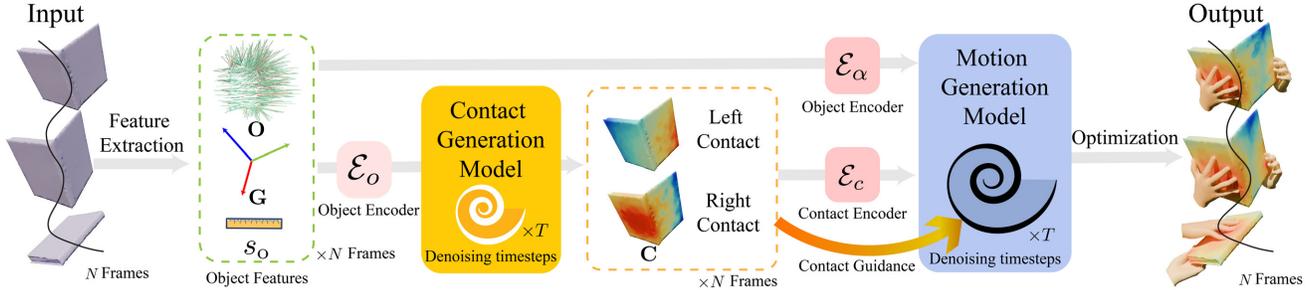
Figure 2. **Overview.** BimArt takes N frames of object trajectories as input and generates N frames of 3D bimanual interactions. The object features (articulation-aware BPS features $\mathbf{O}$, 6D global states $\mathbf{G}$, and the object scale s_o) are passed into both the object encoder $\mathcal{E}_o$ (MLP) in the contact generation model and $\mathcal{E}_\alpha$ (MLP) in the motion generation model. Additionally, the motion generation model's contact encoder $\mathcal{E}_c$ takes $\mathbf{C}$, the bimanual contact map produced by the contact generation model, as conditioning input. The contact model and motion model are both denoising diffusion models, and the spiral denotes the denoising process. $\mathbf{C}$ is further used as guidance at each diffusion timestep to align hand motions with the generated contact maps. Finally, we use optimization to correct contact and penetration artifacts and obtain 3D bimanual meshes.

of seven degrees of freedom: six degrees for the root's orientation and translation, and one degree for the rotational joint. The input to our method is the articulated object trajectory, $\xi = \{\xi_i\}_{i=1}^N, \xi_i = [\mathbf{g}_i | \mathbf{a}_i]$, where $\mathbf{g}_i \in \mathbb{R}^6$ denotes the object's orientation and global translation, and $\mathbf{a}_i \in \mathbb{R}$ represents the articulation angle between the two parts of the object. Given ξ, BimArt generates a corresponding, N-frame bimanual motion $\mathbf{\Theta} = \{\mathbf{\Theta}_i\}_{i=1}^N$, where $\mathbf{\Theta}_i \in \mathbb{R}^{61\times2}$ corresponds to MANO [51] hand parameters for both hands.

Fig. 2 outlines our method. We first introduce an articulation-aware canonicalized feature representation for the object (Sec. 3.1). By keeping the canonicalized object at the origin of the coordinate system, we provide a consistent frame of reference for the object as well as the hands. Next, motivated by the observation that contact understanding facilitates more accurate finger placement, we decompose the task into contact map generation (Sec. 3.2) and motion synthesis based on the generated contact map (Sec. 3.3). Lastly, we use an optimization-based post-processing step to resolve physical artifacts such as penetration and inconsistent contact. (Sec. 3.4).

3.1. Hand and Object Representation

Hand Representation. We encode hand motion in an object-centric way, with each hand at frame i parameterized by both surface keypoint positions $\mathbf{H}_i$ and direction vectors to the object $\mathbf{D}_i$ as shown in Fig. 3. More specifically, $\mathbf{H}_i \in \mathbb{R}^{\mathtt{J}\times3}$ is a sparse set of vertices sampled from the MANO surface vertices $\mathbf{\Xi}_i$. $\mathbf{D}_i \in \mathbb{R}^{\mathtt{J}\times3}$ denotes the direction vectors originating from the hand keypoints $\mathbf{H}_i$ to their nearest object vertices. $\mathbf{D}_i$ encodes both the direction and the magnitude. Compared with the MANO skeletal joints, this representation is denser, making it easier to recover MANO parameters, $\mathbf{\Theta}_i$. In addition, the incorporation of $\mathbf{D}_i$ aids the model in reasoning about contact.
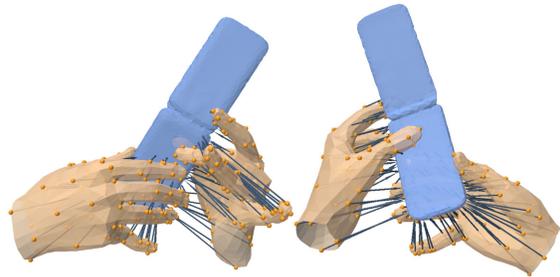
Figure 3. **Hand Representation** We parameterize each frame of hand pose by using J surface keypoints (in orange), sampled from the surface of the hand. In addition to position, we also use the direction vector (dark blue lines) from each keypoint to the nearest object surface as an additional feature.

To generalize to unseen object trajectories and disentangle object motion due to articulation and global trajectory changes, we propose to encode the hand in the object's canonical coordinate frame, *i.e.* the frame where the object's articulation axis is aligned with the negative z-axis. Let $\mathbf{V}_i$ denote the object vertex positions at frame i, and $\mathbf{M}$ be its canonical-to-world transformation matrix. We transform the hand point cloud and object vertices from the world frame $\mathbf{H}_i^w$ to the object's canonical frame $\mathbf{H}_i^o$ such that $\mathbf{H}_i^o = (\mathbf{M})^{-1}\mathbf{H}_i^w$ and $\mathbf{V}_i^o = (\mathbf{M})^{-1}\mathbf{V}_i$. In the rest of the paper, we omit o, since all hand motions are generated in the object's canonical frame.

Object Representation. Next, we define the object feature representation. Training a single model across multiple object types necessitates a feature representation that encodes geometric information consistently while remaining independent of the object topology. We therefore represent the object trajectory using Basis Point Sets (BPS) [50].

The BPS representation requires defining a fixed set of basis points $\mathbf{B} \in \mathbb{R}^{\mathtt{K}\times3}$ that are typically uniformly sampled from the unit sphere. The BPS features are then computed as a set of vectors from $\mathbf{B}$ to the nearest object vertices. This

3

formulation will lead to a suboptimal sampling strategy for objects with articulated parts at different scales. In contrast to the original BPS formulation, we propose to use normalized part-based BPS features computed in the object-centric frame where the same basis point set is mapped separately to each articulated part. Let s_o denote the object scale, computed by normalizing the maximum distance from the origin to the object vertices ($\mathbf{V}_{ao}$) with an open articulation angle in the canonical space:

$$s_o = \frac{1 - d_{\text{margin}}}{\max_{\mathbf{v} \in \mathbf{V}_{ao}} \|\mathbf{v}\|} \quad (1)$$

A margin d_{margin} is used to prevent the object pointcloud from touching the boundary of the unit ball. To provide a denser mapping from basis points $\mathbf{B}$ to object vertices $\mathbf{V}$, the object is normalized to the unit sphere, using s_o. The normalized, part-based BPS features can then be computed as:

$$\mathbf{O}_i^p = \left[\operatorname*{argmin}_{\mathbf{v} \in \mathbf{V}_i^p} d(\frac{\mathbf{v}}{s_o}, \mathbf{b}) - \mathbf{b} \,, \text{for } \mathbf{b} \in \mathbf{B} \right] \quad (2)$$

$$\mathbf{O} = [\mathbf{O}_i^p \,, \text{for } i \in \{1, 2, \ldots, N\}, p \in \{\text{top}, \text{bottom}\}] \quad (3)$$

In Eq. (2), $d(\cdot)$ is the Euclidean distance between two points and p denotes the part index. Notably, we do not perform a part-based scale normalization, since hand motion $\mathbf{H}$ is encoded in the original scale, and having separate object scales in canonical spaces will increase the difficulty for the model in reasoning about hand object distance and contact.

Alternatively, one could sample the basis points in the original scale of the object without normalizing the object to a unit sphere or using a part-agnostic BPS mapping; the comparison for different sampling strategies is shown in Fig. 4. Our part-based BPS with scale normalization provides a denser mapping to the object, thus forming a more detailed descriptor of the object geometry. Crucially, it ensures that the objects with a small articulating part (eg. the lid of a ketchup bottle) are not under-sampled against the larger base.

Our BPS feature $\mathbf{O}$ is independent of the object's global trajectory, encoding only the object's shape and articulation states. However, without encoding the object's global movement, the generated motion will be physically implausible since it is not aware of the gravity direction and cannot distinguish the object trajectories that require a supporting hand at the bottom. Therefore, we further include the global states $\mathbf{G} = [\mathbf{g}_i]_{i=1}^N$ as a lower dimensional 6D vector per frame, which consists of relative translation to the first frame and the global rotation. Relative translation is used to avoid overfitting and increases robustness to unseen test trajectories. Overall, $\mathbf{O}$ and $\mathbf{G}$ capture the detailed object geometry, articulation movement, and global movement.
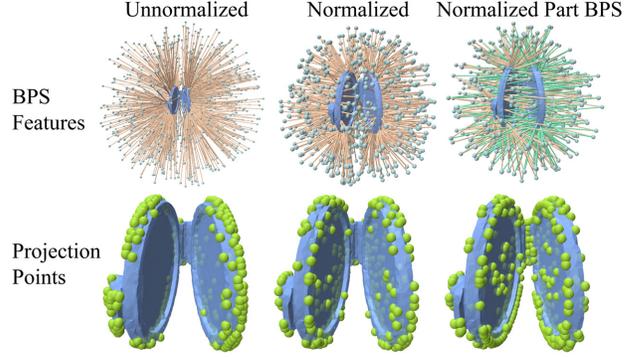
Figure 4. **Different BPS Sampling Strategies**. Top left: $\mathtt{K} \times 2$ basis points sampled uniformly within a 0.5-meter radius for unnormalized objects. Top middle: $\mathtt{K} \times 2$ BPS sampled uniformly in a unit ball for normalized objects. Top right: $\mathtt{K}$ basis points sampled uniformly in a unit ball for normalized objects, with points mapped to each articulated part of the object, maintaining the same feature dimension. Bottom: Green points on the object represent the projections of the BPS feature vectors. The proposed Normalized Part BPS provides denser mapping on the object's inner surface layer.

3.2. Bimanual Contact Generation Model

Having defined our articulated object representation, we introduce our novel denoising diffusion probabilistic model for generating plausible bimanual contact maps. Importantly, our contact model can be jointly trained on *cross-category* articulated objects, thanks to our generalizable object representation.

Given our object BPS features $\mathbf{O}$, global states $\mathbf{G}$, and the object scale s_o, the contact model generates the corresponding sequence of contact maps for the left and right hand, *i.e.* $\mathbf{C} = [\mathbf{C}^\rho], \rho \in \{\text{left}, \text{right}\}$.

Our bimanual contact maps at frame i are defined as the minimum distance from each object vertex $\mathbf{v}$ from any of the hand vertices $\boldsymbol{\Xi}_i^\rho$ for each hand:

$$\mathbf{C}_i^\rho = \left[\operatorname*{argmin}_{\mathbf{h} \in \boldsymbol{\Xi}_i^\rho} d(\mathbf{h}, \mathbf{v}) - \mathbf{v} \,, \text{for } \mathbf{v} \in \tilde{\mathbf{V}}_i \right], \rho \in \{\text{left}, \text{right}\} \quad (4)$$

where $\tilde{\mathbf{V}}_i = \mathbf{O}_i + \mathbf{B}$, the closest object vertices from the basis points $\mathbf{B}$. We generate separate contact maps for left and right hand to reduce ambiguity when using them as guidance in the motion generation (see Sec. 3.3). Note, that our contact map does not encode correspondence between which hand vertex should be in contact with the object vertex, as doing so would over-constrain the sampling process, thereby hindering motion diversity.

For contact generation, we adopt a denoising diffusion probabilistic model [23, 54] with a transformer-encoder architecture [61, 64], trained to directly predict clean samples $\mathbf{C}$. The model's conditioning inputs include our BPS features $\mathbf{O}$, global states $\mathbf{G}$, and s_o, which are processed through an MLP encoder, $\mathcal{E}_o$. We predict a contact value

per BPS feature in $\mathbf{O}$, facilitating cross-object predictions as the output dimension is fixed by the number of BPS points and remains independent of the object's mesh resolution as shown in Eq. (4). Next, we show how these generated contact maps are used to synthesize hand motions (Sec. 3.3).

3.3. Bimanual Hand Motion Model

Given the object features $\mathbf{O}$ and the contact maps $\mathbf{C}$, our motion model generates N frames of hand motions, parameterized by $\mathbf{X} = [\mathbf{H}|\mathbf{D}]$ as illustrated in Fig. 2. Specifically, the conditions include:
- Object Conditioning: The BPS features $\mathbf{O}$, object global states $\mathbf{G}$, and the object scale s_o are encoded using the object encoder $\mathcal{E}_\alpha$ into a latent object embedding $\mathbf{Z}_o \in \mathbb{R}^{N \times L_o}$, where L_o denotes the latent dimension.
- Contact Conditioning: The contact maps $\mathbf{C}$ are encoded using the contact encoder $\mathcal{E}_c$ into a latent contact feature embedding $\mathbf{Z}_c \in \mathbb{R}^{N \times L_c}$.

$\mathcal{E}_\alpha$ and $\mathcal{E}_c$ are MLPs and their respective outputs are concatenated to form $\mathbf{Z} = [\mathbf{Z}_o|\mathbf{Z}_c]$.

Similar to the contact generation model, we use another transformer encoder (denoted as $\mathcal{M}$) as the diffusion denoiser that is also trained to predict the clean samples. To learn smooth and diverse hand motions and counter the potential noise in the contact model's prediction, we train $\mathcal{M}$ using classifier-free guidance by randomly replacing the contact features $\mathbf{C}$ with a learnable null token $\emptyset$ with a probability p_f.

We observe that the latent feature $\mathbf{Z}_c$ effectively guides $\mathcal{M}$ to establish and maintain contact with the articulated object, dynamically adjusting to changing contact patterns while ensuring temporal consistency. However, at a more fine-grained level, the generated motion is not free from physical artifacts such as fingers being stuck between the parts. Therefore, we introduce a contact map discrepancy term during guidance, encouraging the noisy motion at each denoising timestep to more precisely align with the contact map output $\hat{\mathbf{C}}$ from the contact model. Namely, for each predicted clean hand $\hat{\mathbf{H}}_{(t)}^\rho$ at denoising timestep t, we compute a derived contact map $\tilde{\mathbf{C}}_{(t)}^\rho$ from $\hat{\mathbf{H}}_{(t)}^\rho$ to the nearest object vertex:

$$\tilde{\mathbf{C}}_{(t)}^\rho = \left[\underset{\mathbf{h} \in \hat{\mathbf{H}}_{(t)}^\rho}{\arg\min} d(\mathbf{h}, \mathbf{v}) - \mathbf{v} , \text{ for } \mathbf{v} \in \tilde{\mathbf{V}} \right], \rho \in \{\text{left, right}\} \tag{5}$$

In practice, we use a differentiable 1-nearest neighbor function for the above computation to ensure gradient propagation. The contact map guidance can be written as

$$\tilde{\mathbf{X}}_{(t)}^\rho = \hat{\mathbf{X}}_{(t)}^\rho - \lambda_c \nabla_{\hat{\mathbf{X}}_t^\rho} \left\| \hat{\mathbf{C}}^\rho - \tilde{\mathbf{C}}_{(t)}^\rho \right\|, \tag{6}$$

where λ_c is the guidance scale and $\nabla_{\hat{\mathbf{X}}_t^\rho}$ denote the gradient with respect to the discrepancy term. Finally, we ap-

ply classifier-free guidance to combine the outputs with and without $\mathbf{Z}_c$. The predicted clean motion at timestep $t - 1$ can be written as

$$\tilde{\mathbf{X}}_{(t-1)} = (1 + \lambda_f)\tilde{\mathbf{X}}_{(t)} - \mathcal{M}(\hat{\mathbf{X}}^{(t)}, t, \mathbf{Z}_o \emptyset). \tag{7}$$

Since $\mathcal{M}$ does not predict dense MANO surface vertices, we rely on an optimization-based MANO fitting described in the next section to obtain the final 3D bimanual motions.

3.4. Physically Plausible Hand Motion

Our generated hand motions $\hat{\mathbf{H}}$ only contain a subset of MANO surface vertices and, therefore, some hand surface areas may still experience minor penetration, momentary loss of contact, or slight jitter after denoising. To address this, we introduce an optimization-based MANO fitting to further refine the predictions.

First, we estimate the MANO parameters $\boldsymbol{\Theta} = [\boldsymbol{\theta}|\boldsymbol{\beta}]$ for both hands, where $\boldsymbol{\theta} \in \mathbb{R}^{N \times 51 \times 2}$ and $\boldsymbol{\beta} \in \mathbb{R}^{10 \times 2}$, from the predicted $\hat{\mathbf{H}}$. $\boldsymbol{\theta}$ contains the root translation, rotation, and the per-joint rotations of MANO, with all rotations represented as axis-angle vectors. We estimate $\boldsymbol{\Theta}$ by minimizing the following loss:

$$l_{\text{MANO}} = \|\hat{\mathbf{H}} - f_{\text{MANO}}(\boldsymbol{\theta}, \boldsymbol{\beta})\| \tag{8}$$

f_{MANO} is the MANO forward pass operator to retrieve fitted hand keypoints based on optimized $\boldsymbol{\theta}, \boldsymbol{\beta}$.

Next, we refine the estimated MANO parameters $\boldsymbol{\Theta}$ to reduce penetrations, temporal jitter and enforce contact at the predicted points using three energy terms:

$$l_{\text{reg}} = w_{\text{proj}} l_{\text{proj}} + w_{\text{pen}} l_{\text{pen}} + w_{\text{acc}} l_{\text{acc}} \tag{9}$$

Since our denoising outputs contain both $\hat{\mathbf{H}}$ and $\hat{\mathbf{D}}$, the projection loss l_{proj} encourages the projection points of hand keypoints based on the direction vectors, $i.e.$ $\mathbf{P} = f_{\text{MANO}}(\boldsymbol{\theta}, \boldsymbol{\beta}) + \hat{\mathbf{D}}$, to lie on the object surface, resolving the potential floating artifact.

$$l_{\text{proj}} = \sum_{\mathbf{p} \in \mathbf{P}} \min_{\mathbf{v} \in \mathbf{V}} \|\mathbf{p} - \mathbf{v}\| \tag{10}$$

The dense predicted hand surface vertices after MANO fitting, denoted as $\hat{\mathbf{\Xi}}$, are fed into the penetration loss [21]:

$$l_{\text{pen}} = \sum_{\mathbf{h} \in \text{Int}(\hat{\mathbf{\Xi}})} \min_{\mathbf{v} \in \mathbf{V}} \|\mathbf{h} - \mathbf{v}\|, \tag{11}$$

$\text{Int}(\hat{\mathbf{\Xi}})$ refers to the set of hand vertices inside the object. Finally, we penalize the acceleration of hand vertices $\hat{\mathbf{\Xi}}$:

$$l_{\text{acc}} = \sum_{\mathbf{h}_i \in \hat{\mathbf{\Xi}}} \|\mathbf{h}_i - 2 \cdot \mathbf{h}_{i-1} + \mathbf{h}_{i-2}\| \tag{12}$$

w_{proj}, w_{pen} and w_{acc} are hyperparameters. We demonstrate the effectiveness of the optimization in Sec. 4.

5

3.5. Implementation Details

In data preprocessing, d_{margin} is set to 0.15 for scale normalization. The contact model and the motion model share the same architecture hyperparameters, *i.e.* they both have eight transformer encoder layers with latent dimension 512. Both models are trained with 50 diffusion steps on the ARCTIC [14] dataset for 200 epochs, using the Adam optimizer [12] of learning rate $1e^{-4}$ with a cosine learning rate scheduler [42]. DDPM noise schedule [24] is adopted and the models directly predict clean samples. In addition, Exponential Moving Average (EMA) models [22] are used for better stability. The motion model has a contact condition dropout rate of 0.5. For classifier-free guidance [23], the guidance scale λ_f is set to 0.5. We determine the contact map guidance scale λ_c by the gradient norm, *i.e.* $\lambda_c = \frac{1}{\|\nabla\hat{\mathbf{x}}_{(t)}\|}$. Training is completed in less than two days on a single A40 GPU. The post-processing is performed for 100 iterations, with $w_{\text{proj}} = 100, w_{\text{pen}} = 10$ and $w_{\text{acc}} = 1000$ on the ARCTIC dataset. We set w_{acc} to 10000 for the HOI4D dataset [40].

4. Experiments

Datasets. We evaluate our method on the ARCTIC dataset [14], which contains fully annotated mesh sequences for bimanual interactions with 11 articulated objects. For each object, we use four motion sequences as the test set and the rest as training sequences. Among the four sequences, one contains only grasping and object relocation, while the others include both grasping and articulation. In total, we have 257 training sequences and 44 test sequences. In addition, we evaluate our method on HOI4D [40], a large-scale dataset containing 3D annotations of articulated object movements and hand poses. We follow the evaluation protocol of Zheng *et al.* [82] and use two provided categories, pliers and scissors, with the same train and test split.

Evaluation Metrics. Quantitatively measuring synthesized motion has been a challenging pursuit. Toward this goal, we evaluate the methods on a variety of metrics, each measuring a specific aspect of motion generation. The multi-modality metric, which measures the method's ability to generate diverse results for the same object trajectory, is computed using the mean average pairwise distance between all generated hand vertices by sampling 10 times for the same trajectory (denoted as "Mul" in Tab. 1 and Tab. 3). For evaluating the geometric feasibility of the synthesized motions, we evaluate the extent of penetration and contact feasibility. "Pen 1cm" is the percentage of motion frames with hand vertex penetration, using a 1cm threshold. "CM" measures the $l1$ distance of the contact map derived from the generated hand motions from the predicted contact map. This metric is only applicable to our ablations. "Con" mea-

Method	Mul (cm) ↑	Accel $\left(\frac{cm}{s^2}\right)$ ↓	Pen 1cm (%) ↓	Con (%) ↑	Art (%) ↑
GT	-	0.17848	1.0398	95.138	94.563
CAMS-B	8.5602	0.11959	42.519	98.915	76.704
MDM-B	0.55459	0.27666	66.71	93.657	73.734
OMOMO-B	0.038338	0.1969	30.435	96.917	80.094
Ours	6.9093	0.18846	2.0346	99.629	85.572

Table 1. **Quantitative Comparison on ARCTIC.** Our method outperforms the state-of-the-art in penetration, contact, and articulation. Even though CAMS-B scores better in the multimodality and acceleration, it exhibits low interaction plausibility, as seen in high penetration percentage and qualitative results in Fig. 5.

sures the percentage of motion frames with object contact and "Art" measures the percentage of motion frames where the hand is in contact with the articulated part, out of the frames with object articulation changes. In addition, we compute hand vertex penetration percentage, contact, and articulation consistency following CAMS [82]'s protocol on HOI4D.

Baselines. Except for the concurrent work [79], no prior works have tackled bimanual motion synthesis for articulated objects given object trajectories under identical assumptions. Therefore, we propose the following modifications to various baselines:

- CAMS [82] is a category-specific method that produces single-hand motions. CAMS-X denotes the cross-category model trained by us on the HOI4D dataset and CAMS-B is the bimanual model we adapted for the ARCTIC dataset.
- MDM [61] is a pioneer work for diffusion-based motion synthesis. We change text-based conditioning to object trajectory conditioning using our normalized part-based BPS features and denote this variant as MDM-B.
- OMOMO [35] is a whole-body method, which generates human-object interaction without finger articulations. In our adaptation OMOMO-B, we generate hand joints in stage one with contact constraints applied to all the joints. In stage two, we predict the over-parameterized hand motions conditioned on joints.

For more details, we refer to the supplementary document.

		Pliers			Scissors		
		Pen (%) ↓	Con. Score ↑	Art. Score ↑	Pen (%) ↓	Con. Score ↑	Art. Score ↑
	Ground Truth	0.000	1.000	1.000	0.046	1.000	0.970
Cat.Spec.	GraspTTA	0.555	0.779	0.420	0.454	0.993	0.849
	GraspTTA w/ opt	0.294	0.727	0.321	0.812	0.994	0.959
	ManipNet	0.548	0.984	0.892	0.391	0.917	0.417
	ManipNet w/ opt	0.387	0.890	0.738	0.131	0.831	0.333
	CAMS w/ opt	0.563	0.916	0.393	0.590	0.997	0.850
	CAMS	0.004	1.000	1.000	0.080	0.999	0.989
Unified	CAMS-X	0.017	0.485	0.015	0.198	0.858	0.167
	Ours	0.044	0.966	0.597	0.591	1.000	0.853

Table 2. **Evaluation on the HOI4D Dataset.** We show comparisons in the category-specific setting (denoted as "Cat.Spec") and the cross-category setting where a unified model is trained (denoted as "Unified"). The numbers for "Cat.Spec" are taken from CAMS [82]. Our method outperforms CAMS-X, and performs comparatively with methods trained in a category-specific way.
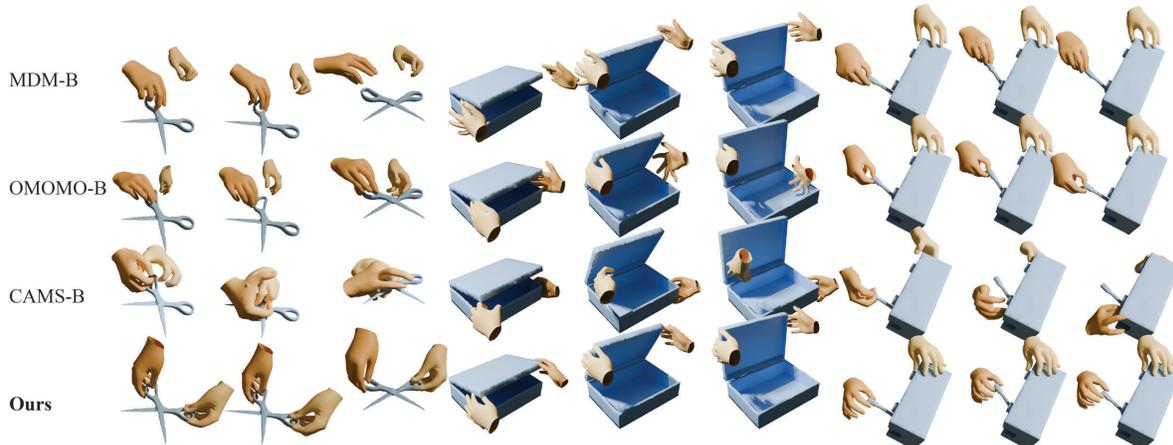
Figure 5. **Qualitative Comparison.** MDM-B struggles with establishing accurate contact, as seen in the hand-object gap in the scissors and the box example. OMOMO-B's rigid contact constraints make it prone to failure, especially with large wrist movements, like opening a box. CAMS-B failed to generate plausible motions, since its stage-wise contact targets under-constrain MANO fitting in dynamic settings with complex contact patterns and diverse object trajectories.

Figure 6. **Diverse Results.** We show diverse bimanual sequences together with the predicted contact maps on the laptop, ketchup and mixer given the same unseen trajectory per object. Our method generates accurate finger placements guided by the predicted contact maps.

	Method	Mul (cm) ↑	Accel ($\frac{cm}{s^2}$) ↓	Pen 1cm (%) ↓	Con (%) ↑	Art (%) ↑	CM(cm) ↓
	GT	-	0.17848	1.0398	95.138	94.563	-
Rep.	U-BPS	6.2551	0.28275	17.542	97.214	85.642	1.1072
	NPA-BPS	6.317	0.28537	17.508	97.442	82.04	1.0789
	MANO-Rep	6.4781	0.27243	22.201	95.255	76.814	1.6379
	NP-BPS w/o **G**	6.4149	0.28233	20.414	98.093	82.079	1.0562
	NP-BPS	6.97928	0.31398	20.273	98.481	83.089	1.1505
Contact.	w/o **C**	4.4894	0.31034	9.4481	96.129	79.591	-
	w **C**	6.97928	0.31398	20.273	98.481	83.089	1.1505
	w **C** + CG	6.9551	0.30371	16.496	97.351	84.227	1.1284
	w C+CG+Opt	6.9093	0.18846	2.0346	99.629	85.572	1.1778

Table 3. Ablations for various object and hand representations and ways to utilize contact information. The experiment in bold is our proposed design.

4.1. Quantitative Results

We tabulate the quantitative comparison of the methods in Tab. 1 and Tab. 2 for the ARCTIC and HOI4D datasets, respectively. Our method outperforms MDM-B and OMOMO-B in all metrics. Even though CAMS-B scores better in multi-modality, acceleration, and contact,

we show qualitatively that CAMS-B struggles to produce natural and plausible motions in Sec. 4.3.

In the single-hand setting on HOI4D (Tab. 2), our cross-category model performs comparatively with the category-specific baselines. In the cross-category setting, we outperform CAMS-X in terms of articulation and contact consistency by a large margin, highlighting the advantage of our method to handle a variety of geometries in a unified manner. We refer the reader to the supplemental video for a holistic assessment of our results.

4.2. Perceptual User Study

The interaction plausibility is difficult to assess using quantitative metrics alone. Hence, we conducted a perceptual user study with 55 human respondents to evaluate our generated motions compared with OMOMO-B and MDM-B. We exclude CAMS-B for the user study, as its motion quality is significantly subpar evident in Fig. 5 and our supple-
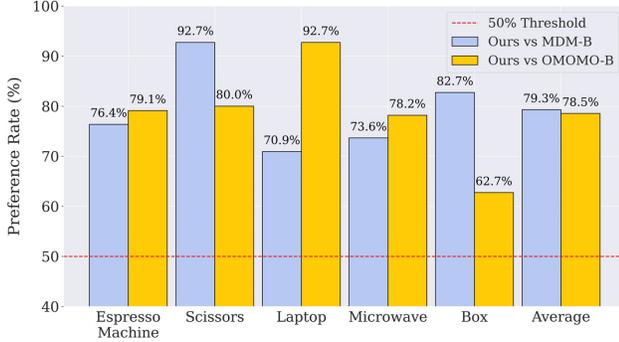
Figure 7. **User Study Results**. We show the preference rate of BimArt against MDM-B and OMOMO-B. Our method outperforms existing state-of-the-art for all objects covered in the user study.

mentary video. The user study contains 40 pairs of animations, covering five objects and four object trajectories randomly sampled from the test set. We split the user study into two subgroups, each covering two out of the four object trajectories with 20 questions. In each question, we present the participants with two animations with the same object trajectory, one of which is generated by our method. The survey has a force-choice style with the following question: *Which animation has a more natural hand motion that aligns better with the object trajectory?* The animations are interactive, allowing the user to zoom or rotate the view to access the quality accurately. We calculate p-values (z-test) for our comparison with each method and observe statistically significant results with $p < 10^{-3}$ for all baselines. We demonstrate that the users prefer our motions for every object in Fig. 7.

4.3. Qualitative Results

We show qualitative comparisons in Fig. 5 and diverse samples from our method in Fig. 6. BimArt generates more natural and physically realistic motions for small objects that require precise contact region geometry understanding, like the "grabbing the scissors" example, and large objects with significant articulation movements, like the "opening box" example in Fig. 5. Fig. 6 shows three samples of hand motions given the same object trajectory, demonstrating BimArt's diversity.

4.4. Ablations

We present ablation studies to investigate the effect of our design choices and report the results in Tab. 3. We split the ablations into two sections, with the "Rep" section ablating the different object representations and hand representations. The BPS representations in Tab. 3 include Unnormalized BPS (U-BPS), normalized part-agonistic BPS (NPA-BPS), and our proposed normalized part-based BPS (NP-BPS), all trained with contact conditions. We also ablate the effect of removing global states **G** from (NP-BPS

w/o **G**), leading to the lack of the object global movement awareness in the contact generation and motion generation model. As an alternative hand representation, MANO-Rep refers to MANO 6D pose parameters with joint positions. Notably, we do not apply contact map guidance and post-refinement in this set of experiments to isolate the effect caused by object and hand representations. The "Contact" section demonstrates the effect of the contact condition, (*i.e.* w/o **C** versus w **C**), contact map guidance (w **C** + CG), and the optimization-based refinement (w **C** + CG + Opt).

The following observations can be drawn from the Tab. 3. NP-BPS leads to the highest multi-modality and contact percentage compared with the alternative BPS sampling strategies. MANO-Rep leads to worse penetration, contact, and articulation percentages, highlighting its limitations compared to our proposed hand representation. NP-BPS w/o G performs better in acceleration and contact map discrepancy by focusing on articulation-aware hand motions, while excluding global states, which we show leads to physically implausible motions in the supplementary video. In contact ablations, contact conditioning leads to a higher multi-modality, contact, and articulation percentage. Contact guidance helps the hand motions better align with the contact maps, evidenced by a lower contact map discrepancy. Our optimization-based refinement significantly reduces the penetration percentage and the hand motion acceleration while maintaining contact with the objects. For qualitative ablations, please refer to the supplementary video.

5. Conclusion

Limitations. Although fairly robust to novel object trajectories, our method is restricted to the limited number of object categories as provided in the datasets (ARCTIC and HOI4D). In the real world, however, one would like to generalize to new (and open-vocabulary) objects in a zero-shot manner. We believe leveraging the common-sense knowledge of existing multi-modal large language models would facilitate such generalization [26, 37]. Our method could also benefit from the incorporation of faster diffusion sampling approaches such as DDIM [55], or Latent Diffusion Modeling [7] to facilitate adoption in artistic creation processes with limited time budgets.

In this work, we proposed BimArt, a novel bimanual motion synthesis method given the trajectory of an articulated 3D object. Our proposed feature representation leads to high diversity in generated motions, providing the flexibility for 3D artists and animators to sample multiple plausible interactions for a single object trajectory. In both quantitative metrics and the user study, our approach outperforms competing methods in terms of naturalness and physical plausibility, paving the way for more realistic and user-friendly hand-object animation.

References

[1] Ijaz Akhter, Tomas Simon, Sohaib Khan, Iain Matthews, and Yaser Sheikh. Bilinear spatiotemporal basis models. *ACM Transactions on Graphics (TOG)*, 2012. 2

[2] Samarth Brahmbhatt, Cusuh Ham, Charles C. Kemp, and James Hays. ContactDB: Analyzing and predicting grasp contact via thermal imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[3] Matthew Brand and Aaron Hertzmann. Style machines. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, 2000. 2

[4] Jona Braun, Sammy Christen, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Physically plausible full-body hand-object interaction synthesis. In *International Conference on 3D Vision (3DV)*, 2024. 2

[5] Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2hoi: Text-guided 3d motion generation for hand-object interaction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[6] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[7] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 8

[8] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems*, 2023. 2

[9] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-Grasp: Physically Plausible Dynamic Grasp Synthesis for Hand-Object Interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 13

[10] Sammy Christen, Shreyas Hampali, Fadime Sener, Edoardo Remelli, Tomas Hodan, Eric Sauser, Shugao Ma, and Bugra Tekin. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. *arXiv preprint arXiv:2403.17827*, 2024. 2

[11] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[12] P Kingma Diederik. Adam: A method for stochastic optimization. *The International Conference on Learning Representations*, 2014. 6

[13] Enes Duran, Muhammed Kocabas, Vasileios Choutas, Zicong Fan, and Michael J. Black. Hmp: Hand motion priors for pose and shape estimation from video. *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024. 2

[14] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 6, 13, 14

[15] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015. 2

[16] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions. In *Computer Graphics Forum (Eurographics)*, 2023. 1, 2, 13

[17] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Remos: 3d motion-conditioned reaction synthesis for two-person interactions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2

[18] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[19] Yuze Hao, Jianrong Zhang, Tao Zhuo, Fuan Wen, and Hehe Fan. Hand-centric motion refinement for 3d hand-object interaction via hierarchical spatial-temporal modeling. *Association for the Advancement of Artificial Intelligence*, 2024. 2

[20] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1

[21] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 6

[23] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 4, 6

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 2020. 6

[25] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)*, 2017. 2

[26] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Inject-

ing the 3d world into large language models. *Advances in Neural Information Processing Systems*, 2023. 8

[27] Binghao Huang, Yuanpei Chen, Tianyu Wang, Yuzhe Qin, Yaodong Yang, Nikolay Atanasov, and Xiaolong Wang. Dynamic handover: Throw and catch with bimanual hands. *Conference on Robot Learning*, 2023. 2

[28] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013. 2

[29] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[30] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *International Conference on 3D Vision (3DV)*, 2020. 2

[31] Korrawe Karunratanakul, Konpat Preechakul, Emre Aksan, Thabo Beeler, Supasorn Suwajanakorn, and Siyu Tang. Optimizing diffusion noise can serve as universal motion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[32] Jihyun Lee, Shunsuke Saito, Giljoo Nam, Minhyuk Sung, and Tae-Kyun Kim. Interhandgen: Two-hand interaction generation via cascaded reverse diffusion. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[33] Kang-Won Lee, Yuzhe Qin, Xiaolong Wang, and Soo-Chul Lim. Dextouch: Learning to seek and manipulate objects with tactile dexterity. *arXiv preprint arXiv:2401.12496*, 2024. 2

[34] Andreas M Lehrmann, Peter V Gehler, and Sebastian Nowozin. Efficient nonlinear markov models for human motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[35] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 2023. 1, 2, 6, 13

[36] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2

[37] Lei Li and Angela Dai. Genzi: Zero-shot 3d human-scene interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 8

[38] Shaowei Liu, Yang Zhou, Jimei Yang, Saurabh Gupta, and Shenlong Wang. Contactgen: Generative contact modeling for grasp generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[39] Xueyi Liu and Li Yi. Geneoh diffusion: Towards generalizable hand-object interaction denoising via denoising diffusion. In *Int. Conf. Learn. Represent.*, 2024. 1, 2

[40] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 6, 13

[41] Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[42] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *The International Conference on Learning Representations*, 2017. 6

[43] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[44] Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[45] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[46] Muhammad Hamza Mughal, Rishabh Dabral, Ikhsanul Habibie, Lucia Donatelli, Marc Habermann, and Christian Theobalt. Convofusion: Multi-modal conversational diffusion for co-speech gesture synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[47] Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. *arXiv preprint arXiv:2312.06553*, 2023. 2

[48] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[49] Mathis Petrovich, Or Litany, Umar Iqbal, Michael J. Black, Gül Varol, Xue Bin Peng, and Davis Rempe. STMC: Multi-track timeline control for text-driven 3d human motion generation. *CVPR Workshop on Human Motion Generation*, 2024. 2

[50] Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2, 3

[51] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 2017. 3

[52] Nicholas Sharp, Yousuf Soliman, and Keenan Crane. The vector heat method. *ACM Transactions on Graphics (TOG)*, 2019. 13, 15

[53] Soshi Shimada, Franziska Mueller, Jan Bednarik, Bardia Doosti, Bernd Bickel, Danhang Tang, Vladislav Golyanik, Jonathan Taylor, Christian Theobalt, and Thabo Beeler. Macs: Mass conditioned 3d hand and object motion synthesis. In *International Conference on 3D Vision (3DV)*, 2024. 2, 13

[54] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2015. 4

[55] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Int. Conf. Learn. Represent.*, 2021. 8

[56] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Transactions on Graphics (TOG)*, 2019. 1

[57] Sebastian Starke, Ian Mason, and Taku Komura. Deepphase: periodic autoencoders for learning motion phase manifolds. *ACM Transactions on Graphics (TOG)*, 2022. 2

[58] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[59] Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. GOAL: Generating 4D whole-body motion for hand-object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 13

[60] Purva Tendulkar, Dídac Surís, and Carl Vondrick. Flex: Full-body grasping without full-body grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[61] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *Int. Conf. Learn. Represent.*, 2023. 2, 4, 6

[62] Dylan Turpin, Liquan Wang, Eric Heiden, Yun-Chun Chen, Miles Macklin, Stavros Tsogkas, Sven Dickinson, and Animesh Garg. Grasp'd: Differentiable contact-rich grasp synthesis for multi-fingered hands. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[63] Dylan Turpin, Tao Zhong, Shutong Zhang, Guanglei Zhu, Eric Heiden, Miles Macklin, Stavros Tsogkas, Sven Dickinson, and Animesh Garg. Fast-grasp'd: Dexterous multi-finger grasp generation through differentiable simulation. In *International Conference on Robotics and Automation]*, 2023. 2

[64] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 4

[65] Weikang Wan, Haoran Geng, Yun Liu, Zikang Shan, Yaodong Yang, Li Yi, and He Wang. Unidexgrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[66] Jun Wang, Yuzhe Qin, Kaiming Kuang, Yigit Korkmaz, Akhilan Gurumoorthy, Hao Su, and Xiaolong Wang. CyberDemo: Augmenting Simulated Human Demonstration for Real-World Dexterous Manipulation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[67] Jack M. Wang, David J. Fleet, and Aaron Hertzmann. Gaussian process dynamical models for human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008. 2

[68] Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. *International Conference on Robotics and Automation]*, 2023. 2

[69] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1, 2

[70] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[71] Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[72] Yufei Ye, Abhinav Gupta, Kris Kitani, and Shubham Tulsiani. G-hop: Generative hand-object prior for interaction reconstruction and grasp synthesis. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[73] Hongwei Yi, Justus Thies, Michael J. Black, Xue Bin Peng, and Davis Rempe. Generating human interaction motions in scenes with text control. *arXiv:2404.10685*, 2024. 2

[74] Ying Yuan, Haichuan Che, Yuzhe Qin, Binghao Huang, Zhao-Heng Yin, Kang-Won Lee, Yi Wu, Soo-Chul Lim, and Xiaolong Wang. Robot synesthesia: In-hand manipulation with visuotactile sensing. *International Conference on Robotics and Automation]*, 2024. 2

[75] Chenyangguang Zhang, Yan Di, Ruida Zhang, Guangyao Zhai, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Ddf-ho: Hand-held object reconstruction via conditional directed distance field. *Adv. Neural Inform. Process. Syst.*, 36, 2024. 2

[76] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. ManipNet: Neural manipulation synthesis with a Hand-Object spatial representation. *ACM Transactions on Graphics (TOG)*, 2021. 2, 13

[77] Hui Zhang, Sammy Christen, Zicong Fan, Otmar Hilliges, and Jie Song. GraspXL: Generating grasping motions for diverse objects at scale. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 1

[78] Hui Zhang, Sammy Christen, Zicong Fan, Luocheng Zheng, Jemin Hwangbo, Jie Song, and Otmar Hilliges. ArtiGrasp: Physically plausible synthesis of bi-manual dexterous grasping and articulation. In *International Conference on 3D Vision (3DV)*, 2024. 1, 2, 13

11

[79] Jiajun Zhang, Yuxiang Zhang, Liang An, Mengcheng Li, Hongwen Zhang, Zonghai Hu, and Yebin Liu. Manidext: Hand-object manipulation synthesis via continuous correspondence embeddings and residual-guided diffusion, 2024. 6

[80] Wanyue Zhang, Rishabh Dabral, Thomas Leimkühler, Vladislav Golyanik, Marc Habermann, and Christian Theobalt. Roam: Robust and object-aware motion generation using neural pose descriptors. *International Conference on 3D Vision (3DV)*, 2024. 2

[81] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1

[82] Juntian Zheng, Qingyuan Zheng, Lixing Fang, Yun Liu, and Li Yi. Cams: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 6, 13

[83] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object-to-hand correspondence for motion refinement. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[84] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Gears: Local geometry-aware hand-object interaction synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[85] Zehao Zhu, Jiashun Wang, Yuzhe Qin, Deqing Sun, Varun Jampani, and Xiaolong Wang. Contactart: Learning 3d interaction priors for category-level articulated object and hand poses estimation. *International Conference on 3D Vision (3DV)*, 2024. 2

A. Appendix – Overview

In this document, we first present a conceptual comparison of our setting and approach in Appendix B. We provide additional details such as data processing (Appendix C), baseline adaptation (Appendix D), and additional results (Appendix E). Please refer to the supplementary video for animations.

B. Conceptual Comparison

Our method relies on fewer assumptions than the prior works as shown in Tab. 4.

C. Data Processing

We follow the convention of the ARCTIC dataset [14], defining the canonical space as the configuration where the articulation axis aligns with the negative z-axis. For scale normalization, we apply a heuristic to determine an articulation angle that positions the object at a state likely to maximize its distance from the origin. Specifically, we set the articulation angle to $\frac{\pi}{2}$. For the mixer and capsule machine, and to 0 for the scissors and espresso machine. For all other objects, we set the articulation angle to π.

D. Baseline Details

OMOMO Adaptation In the original full-body setting, OMOMO [35] predicts only the wrist positions in stage one. Since the OMOMO dataset lacks finger articulation data, the wrist, being the closest joint to the object, is the natural choice for applying contact constraints. In contrast, in our hand-only setting, all joints have the potential to interact with objects. Limiting contact constraints to the wrist in this context would be suboptimal. Therefore, we design stage one to predict all hand joints, applying contact constraints to each joint. In stage two, we refine the motion predictions by estimating the hand poses, conditioned on all joints.

ArtiGrasp We also re-trained ArtiGrasp [78] on our train/test split and evaluated the dynamic object grasping and articulation task which performs grasping and articulation in separate stages. Since the object's initial state has to be supported by the table in the simulator, we set the relative change of the object state to be the same without violating the physical constraint (eg. the goal state should not penetrate the table). ArtiGrasp cannot reach the object goal state reliably at every run, unavoidably, the actual object trajectory from the physics simulator will deviate significantly from ours. Moreover, ArtiGrasp employs heuristics transitioning from grasping to articulation, such as dropping the object on the table and moving the hands apart before articulating, resulting in low contact and articulation percentage. Due to the difficulty in standardizing the setting,

Method	Articulated Objects	Bimanual	No Grasp Ref.	Unified
ManipNet [76]	✗	✓	✓	✓
GOAL [59]	✗	✓	✓	✓
IMOS [16]	✗	✓	✗	✓
MACS [53]	✗	✓	✓	(✗)
D-Grasp [9]	✗	✗	✗	✓
ArtiGrasp [78]	✓	✓	✗	✓
CAMS [82]	✓	✗	✗	✗
BimArt	✓	✓	✓	✓

Table 4. **Conceptual Comparison to Prior Works.** We highlight that our work is the only one, which provides all desired functionalities. **No Grasp Ref.** means that neither initial pose nor goal pose are given as input. **Unified** refers to a single model that can handle various object categories. MACS is only trained on spheres, hence a bracket is added for the checkmark under Unified.

we exclude ArtiGrasp from our quantitative and qualitative comparisons.

E. Additional Results

Besides providing the penetration percentage at 1cm threshold in the main paper, we additionally provide it at 5mm as shown in Tab. 5.

Method	Pen 5mm (%) ↓
GT	30.4
CAMS-B	87.5
MDM-B	66.7
OMOMO-B	74.9
Ours	32.8

Table 5. Penetration percentage at 5mm threshold

Qualitatively, we visualize diverse contact maps our method generates in Fig. 9. Fig. 8 shows the generalization ability of our method to intra-class variations in the HOI4D dataset [40]. Our model is trained in a cross-category manner and we show the qualitative results for all six unseen objects.

F. BPS Analysis

We present additional BPS feature analysis in Tab. 6, by interpolating the contact values associated with sparse object vertices mapped by the basis points using [52] and compute the L1 loss for the densified per vertex contact maps and the ground truth contact maps. A lower error reflects a denser BPS mapping and better geometric representation. The results are broken down into cross-category averages and object-specific errors, with errors reported for

	Average	Microwave	Phone	Box	Ketchup	Mixer	Waffle Iron	Capsule Machine	Notebook	Scissors	Laptop	Espresso Machine
U-BPS-Top	0.546	0.608	0.244	0.454	0.387	0.838	0.705	0.513	0.589	0.401	0.484	0.746
PA-BPS-Top	0.342	0.507	0.137	0.415	0.221	0.377	0.427	0.394	0.288	0.093	0.341	0.533
P-BPS-Top	0.258	0.327	0.152	0.373	0.114	0.336	0.413	0.185	0.265	0.081	0.349	0.216
U-BPS-Bottom	0.552	0.651	0.25	0.5	0.387	0.725	0.57	0.572	0.543	0.543	0.523	0.809
PA-BPS-Bottom	0.341	0.482	0.466	0.194	0.377	0.349	0.103	0.36	0.378	0.374	0.263	0.145
P-BPS-Bottom	0.38	0.645	0.173	0.507	0.232	0.46	0.368	0.472	0.27	0.094	0.395	0.536
U-BPS	0.554	0.645	0.247	0.48	0.387	0.763	0.643	0.568	0.562	0.468	0.504	0.807
PA-BPS	0.32	0.487	0.14	0.444	0.199	0.366	0.404	0.35	0.272	0.099	0.36	0.378
P-BPS	0.361	0.603	0.163	0.449	0.208	0.418	0.393	0.453	0.268	0.087	0.373	0.527

Table 6. **Contact Map Error (in cm) due to BPS mapping.** We present the average and per-category contact map errors resulting from the sparse mapping of BPS features. Both part-agnostic BPS (PA-BPS) and the proposed part BPS (P-BPS) achieve a denser mapping compared to BPS features without scale normalization (U-BPS), resulting in smaller contact map errors. The proposed part-based BPS method further enhances mapping density for the top part of the object (which corresponds to the movable part in canonical space), by allocating equal feature dimensions to individual parts irrespective of their surface area.

Figure 8. **Qualitative Results on HOI4D**. We present visualizations of results for six unseen objects from the HOI4D dataset. Each row illustrates three frames corresponding to the actions of approaching, lifting, and articulating. Notably, our model is trained in a cross-category way.

agnostic BPS (PA-BPS) and the proposed part-based BPS (P-BPS) achieve lower contact errors compared to unnormalized BPS (U-BPS) with the same BPS feature dimensions. PA-BPS achieves a lower average contact map error for the object's bottom parts as the they tend to have a larger surface area in the ARCTIC dataset [14]. Notably, P-BPS reduces the contact map error for the objects' top parts (the movable component in our canonical space) by allocating equal feature dimensions to the top and bottom parts.

the top part, bottom part, and whole object. Both part-
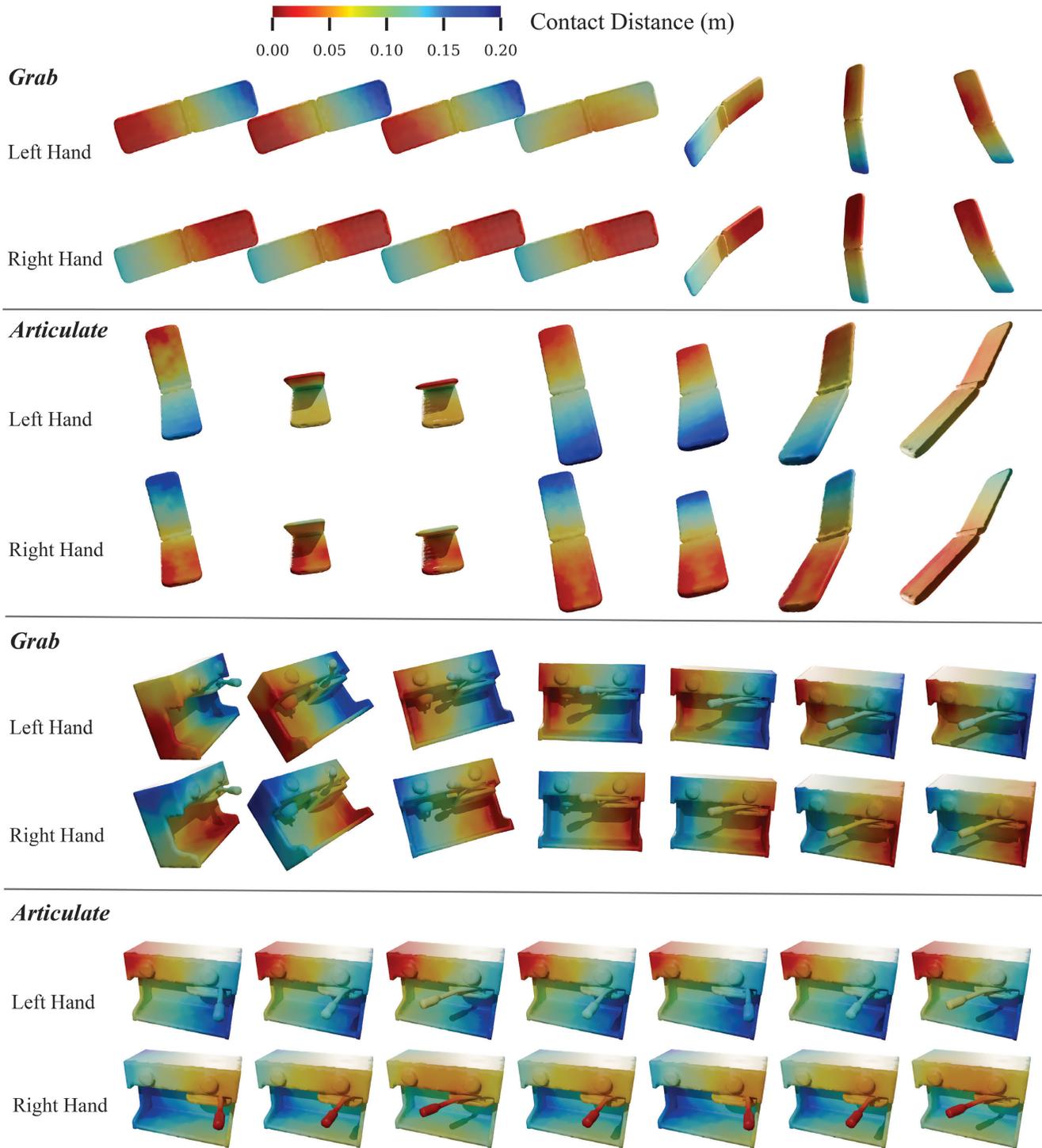
Figure 9. **Contact map visualizations** We present visualizations of the predicted left and right contact maps for seven frames in a sequence. For each object, we include two examples: a "grab" scenario, where the object's articulation remains unchanged, and an "articulate" scenario, where the object undergoes articulation. In the "articulate" examples, the contact region is established at the moving part and remains consistent throughout the articulation process. In contrast, the "grab" examples reveal shifts in the grasping patterns, suggesting that one hand holds the object while the other adjusts its contact point. The Vector Heat method [52] is employed to interpolate the contact values from the sampled object vertices to the full object surface. The predicted contact values are then normalized to a range between 0 and 0.2 meters. In the resulting visualization, red indicates that the hand should be close to the object's surface, while blue signifies that the hand is farther away.