

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

Supplementary Material

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

A. Conceptual Comparison

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

B. 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 .

C. 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 [79] 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

Method	Articulated Objects	Bimanual	No Grasp Ref.	Unified
ManipNet [77]	✗	✓	✓	✓
GOAL [60]	✗	✓	✓	✓
IMOS [16]	✗	✓	✗	✓
MACS [54]	✗	✓	✓	(✗)
D-Grasp [9]	✗	✗	✗	✓
ArtiGrasp [79]	✓	✓	✗	✓
CAMS [83]	✓	✗	✗	✗
BimArt	✓	✓	✓	✓

Table I. **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.

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, we exclude ArtiGrasp from our quantitative and qualitative comparisons.

D. Additional Results

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

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

Table II. Penetration percentage at the 5mm threshold

To show that we are not overfitting to the ground truth, we compute the five nearest neighbors in the training set for each test sequence based on object motions, with the first frame of object vertices centered at zero. We obtain a 15.08 cm average hand vertex distance with a 4.40cm average object vertex distance, showing that our generated mo-

	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 III. **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.

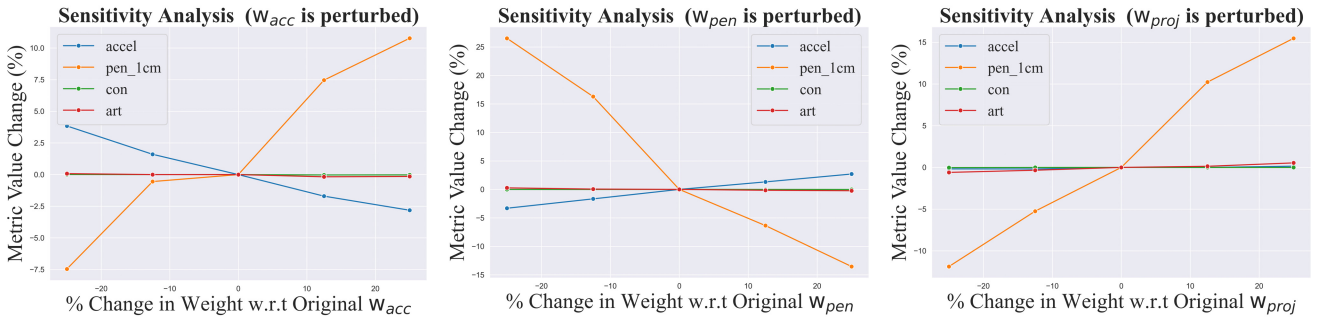


Figure I. Sensitivity analysis for w_{acc} (left plot), w_{pen} (middle plot) and w_{proj} (right plot). We perturb each hyperparameter within $\pm 25\%$ and report the changes in the acceleration, penetration, contact, and articulation metrics.

tions differ from the training ground truth. Please see the supplementary video for qualitative results.

In Fig. I, we show sensitivity analysis plots for w_{acc} , w_{proj} and w_{pen} respectively by perturbing each hyperparameter by $\pm 25\%$ of its original weight. We show the percentage change in the acceleration, articulation, contact, and penetration metrics for each plot. We observe that contact and articulation are not very sensitive to the hyperparameter perturbations, and there exists a trade-off between assigning a higher weight for w_{acc} and assigning a higher weight for w_{pen} as evident in the first 2 plots in Fig. I. A higher weight for w_{acc} leads to better motion smoothness but it increases penetration, and vice versa, when we increase w_{pen} , the motion gets more jittery.

Qualitatively, we visualize diverse contact maps our method generates in Fig. III. Fig. II 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.

E. BPS Analysis

We present additional BPS feature analysis in Tab. III, by interpolating the contact values associated with sparse object vertices mapped by the basis points using [53] 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 the top part, bottom part, and whole object. Both part-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 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.



Figure II. **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.

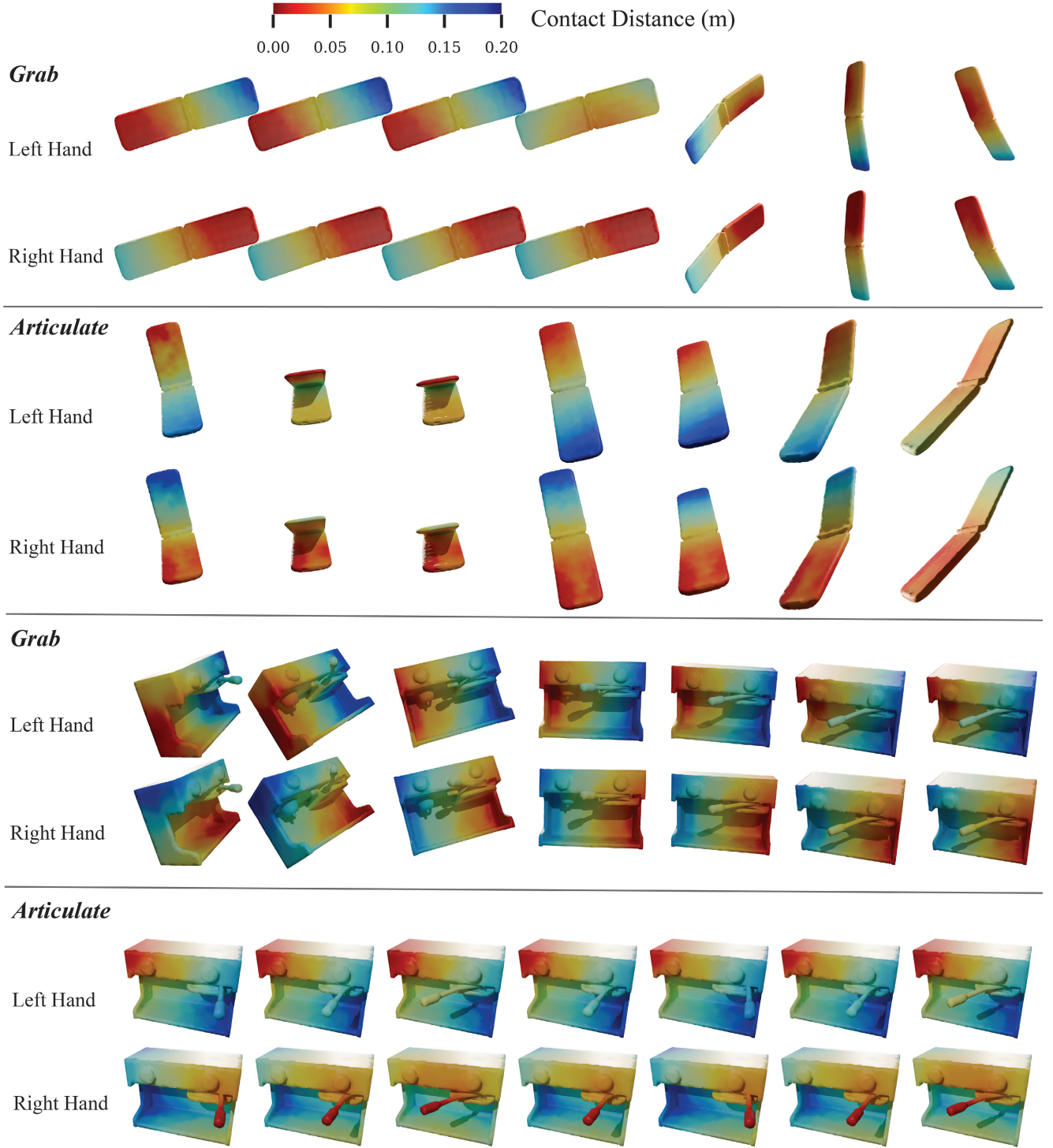


Figure III. **Contact map visualizations.** We present visualizations of the predicted left and right contact maps for seven frames in a sequence. For each object, we include two examples: a “grab” scenario, where the object’s articulation remains unchanged, and an “articulate” scenario, where the object 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 [53] 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.