BiAssemble: Learning Collaborative Affordance for Bimanual Geometric Assembly

Yan Shen*, Ruihai Wu*, Yubin Ke, Xinyuan Song, Zeyi Li, Xiaoqi Li, Hongwei Fan, Haoran Lu, Hao Dong

***Abstract*— Shape assembly, the process of combining parts into a complete whole, is a crucial robotic skill with broad real-world applications. Among various assembly tasks, geometric assembly—where broken parts are reassembled into their original form (*e.g.*, reconstructing a shattered bowl)—is particularly challenging. This requires the robot to recognize geometric cues for grasping, assembly, and subsequent bimanual collaborative manipulation on varied fragments. In this paper, we exploit the geometric generalization of point-level affordance, learning affordance aware of bimanual collaboration in geometric assembly with long-horizon action sequences. To address the evaluation ambiguity caused by geometry diversity of broken parts, we introduce a real-world benchmark featuring geometric variety and global reproducibility. Extensive experiments demonstrate the superiority of our approach over both previous affordance-based and imitation-based methods. Project page: https://sites.google.com/view/biassembly/.**

I. Introduction

Shape assembly, the task of assembling individual parts into a complete whole, is a critical skill for robots with wide-ranging real-world applications. This task can be broadly categorized into two main branches: furniture assembly [1], [2], [3] and geometric assembly [4], [5], [6]. Furniture assembly focuses on combining functional components, such as chair legs and arms, into a fully constructed piece, emphasizing both the functional role of each part and the overall structural design. In contrast, geometric assembly involves reconstructing broken objects, like piecing together parts of a shattered mug, to restore their original form. While furniture assembly has been relatively well-studied—ranging from computer vision tasks that predict part poses in the assembled object [1] to robotic systems that assemble parts in both simulation [7], [8], [9] and real-world environments [2], [10], [11]—geometric assembly remains under-explored despite its significant potential for real-world applications [5], [12], such as repairing broken household items, reconstructing archaeological artifacts [13], assembling irregularly shaped objects in industrial tasks, aligning bone fragments in surgery [14], and reconstructing fossils in paleontology [15].

Previous works on geometric assembly primarily focused on generating physically plausible broken parts through precise physics simulations in the graphics domain [5], [16], and estimating the target assembled part poses based on observations in the computer vision domain [4], [6], [17]. These studies only consider the geometries and ideal assembled poses of broken parts, dismissing the process of step-by-step assembling parts to the complete shape. However, different from opening a door or closing a drawer, only with the ideal part poses, it is difficult for a robot to directly and successfully manipulate broken parts to the complete shape.

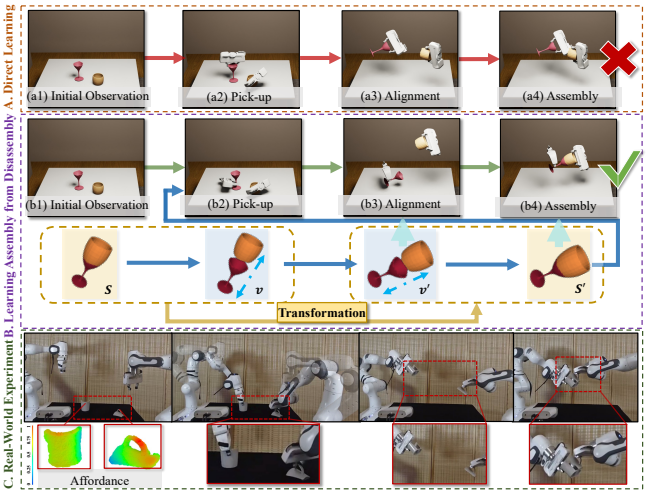

Fig. 1. **(A)** Direct learning long-horizon action trajectories of geometric assembly may face many challenges: grasping ungraspable points, grasping points not suitable for assembly (*e.g.*, seams of fragments), robot colliding with parts and the other robot. **(B)** We formulate this task into 3 steps: pick-up, alignment and assembly. For assembly, we predict the direction that will not result in part collisions. For alignment, we transformed any assembled poses to poses easy for the robot to manipulate from the initial poses without collisions. For pick-up, we learn point-level affordance aware of graspness and the following 2 steps. **(C)** Real-World Evaluations with affordance predictions on two mugs and the corresponding manipulation.

The challenges of the above robotic geometric shape assembly task mainly come from the exceptionally large observation and action spaces. For the observation space, the broken parts have arbitrary geometries, and the graspness on the object surface should consider not only the local geometry itself, but also whether grasping on such point can afford the subsequent bimanual assembly actions. For the action space, as illustrated in Figure 1, it requires long-horizon action trajectories. Given the contact-rich nature of the task, where collisions among the two parts and two robots will easily exist, the actions should be fine-grained and aware of bimanual collaboration. Consequently, the policy must account for geometry, contact-rich assembly processes, and bimanual coordination.

We propose our **BiAssemble** framework for this challenging task. For geometric awareness, we utilize point-level affordance, which is trained to focus on local geometry. This approach has demonstrated strong geometric generalization in diverse tasks [18], [19], including short-term bimanual manipulation [20], such as pushing a box or lifting a basket. To enhance the affordance model with an understanding of subsequent long-horizon bimanual assembly actions, we draw inspiration from how humans intuitively assemble fragments: after picking up two fragments, we align them at the

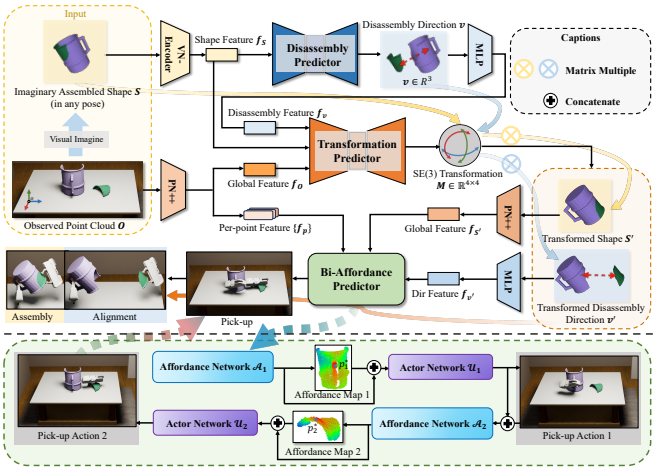

Fig. 2. **Framework Overview.** With the point cloud observation and Imaginary Assembled Shape, the model predicts the disassembly direction in which the disassembled part poses can be easily reached by manipulating the raw parts under the guidance Bi-Affordance.

seam, deliberately leaving a gap (since directly placing them in the target pose often causes geometric collisions), with part poses denoted as alignment poses. We then gradually move the fragments toward each other to fit them together precisely. The alignment poses of the two fragments can be obtained by disassembling assembled parts in opposite directions. With this information, it becomes straightforward to extend the geometry-aware affordance to further be aware of whether the controller can move fragments into the alignment poses without collisions.

We develop a simulation environment where robots can be controlled to assemble broken parts. This simulation environment bridges the gap between vision-based pose prediction for broken parts and the real-world robotic geometric assembly. Moreover, since broken parts exhibit varied geometries (*e.g.*, the same bowl falling from different heights breaking into different groups of fragments), it is challenging to fairly assess policy performance in real-world settings. To address this, we further introduce a real-world benchmark featuring globally available objects with reproducible broken parts, along with their corresponding 3D meshes, which can be integrated into simulation environment. This benchmark enables consistent and fair evaluation of robotic geometric assembly policies. Extensive experiments on diverse categories demonstrate the superiority of our method.

II. Method

A. Overview

Our BiAssembly framework is designed to predict collaborative affordance and gripper actions for bimanual geometric shape assembly. First, to propose the assembly direction on two aligned parts, we develop the Disassembly Predictor to learn the feasible disassembly directions in which the opposite assembly direction will result in no collisions, based on the fracture geometry of the imaginary assembled shape in any pose (II-B). Next, we design the Transformation Predictor, to transform disaassembled parts to poses where the controller can successfully manipulate the initial parts to these alignment poses (II-C). Based on the predicted part alignment poses, we propose the BiAffordance Predictor, which not only predicts where to grasp the fractured parts, but also considers the subsequent collaborative alignment and assembly steps (II-D).

B. Disassembly Prediction Based on Shape Geometry

Feasible disassembly directions (in which the disassembly and opposite assembly processes will not result in collisions) are determined by the fracture geometry of part pairs. We predict these directions from an object-centric perspective on the imaginary assembled shape S in any pose. Notably, these disassembly directions exhibit SO(3) equivariance: they rotate consistently with the parts, allowing separation of shape geometry from pose [4]. To capture this property, we use VN-DGCNN [4] to encode S and obtain an SO(3)-equivariant feature f_s. The Disassembly Predictor, implemented as a conditional variational autoencoder (cVAE), conditions on f_s to model the distribution over disassembly directions.

C. Transformation Prediction For Alignment Pose

Given a collision-free disassembly direction, we predict alignment poses that allow the robot to move parts from initial poses to pre-assembly poses without collisions. This is framed as predicting an SE(3) transformation $M \in \mathbb{R}^{4\times4}$ applied to the assembled shape S and disassembly direction v. PointNet++ encodes initial point cloud O into a global feature f_O, and an MLP encodes v into f_v. A cVAE then takes (f_O, f_v) and predicts M. Applying M to S and v yields transformed S' and v' for alignment and assembly.

D. BiAffordance Predictor

The BiAffordance Predictor proposes bimanual grasps during pick-up step, identifying easy-to-grasp regions while avoiding seams and potential collisions in future alignment and assembly. Following DualAfford [20], we decompose the task into two conditional predictions: the first Affordance and Actor Networks select a grasp point and orientation for one gripper $g_1 = (p_1^*, r_1)$, and the second pair predicts $g_2 = (p_2^*, r_2)$ conditioned on the first. Unlike prior work focused on short-term tasks, our model ensures grasps support downstream alignment and assembly. PointNet++ encodes the initial observation O and transformed shape S', an MLP encodes the transformed disassembly direction v'. The Affordance Networks (MLP) predict per-point graspability, and Actor Networks (cVAE) predict gripper orientations.

E. Alignment and Assembly Actions

After grasping parts, we predict gripper poses for alignment g_i^{align} and assembly g_i^{asm}. We assume gripper-object relative pose remains constant throughout manipulation: $g_i^{pick} \cdot q_i^{pick} = g_i^{asm} \cdot q_i^{asm}; g, q \in SE(3)$. This allows us to compute g_i^{asm} using the known pick-up pose g_i^{pick}, where q_i^{pick} is obtained via a pretrained model[21], and the target part assembled pose is computed by applying the predicted M to the initial pose, $q_i^{asm} = M \cdot q_i^{init}$. The final gripper pose g_i^{asm} is then given by: $g_i^{asm} = g_i^{pick} \cdot q_i^{pick} \cdot (q_i^{init})^{-1} \cdot M^{-1}$.

TABLE I

QUANTITATIVE RESULTS FOR NOVEL INSTANCES WITHIN TRAINING CATEGORIES AND FOR UNSEEN CATEGORIES.

Method	Novel Instances in Training Categories											Novel Categories					
											AVG						AVG
ACT	2%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0.30%	0%	1%	0%	0%	1%	0.4%
Heuristic	5%	8%	0%	3%	2%	4%	3%	5%	10%	2%	4.20%	1%	5%	2%	0%	14%	4.4%
SE(3)-Equiv	0%	0%	4%	0%	1%	5%	0%	7%	2%	11%	3.00%	4%	0%	2%	0%	2%	1.6%
DualAfford	21%	17%	0%	2%	2%	4%	14%	8%	10%	6%	8.40%	5%	10%	4%	1%	16%	7.2%
Ours	**60%**	**38%**	**13%**	**13%**	**12%**	**9%**	**26%**	**18%**	**27%**	**25%**	**24.10%**	**14%**	**31%**	**10%**	**7%**	**25%**	**17.4%**

Notably, this method avoids needing the absolute value of q_i^{init}, relying only on relative transformations. A similar formulation applies to the alignment step, with an additional offset v' added to the transformation.

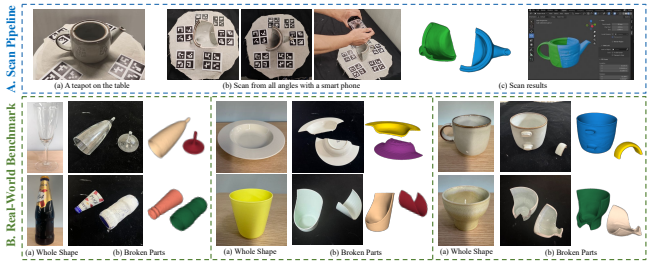

Fig. 3. **Part A** illustrates the pipeline for scanning and reconstructing real objects. **Part B** presents examples of fractured parts from various categories.

III. BENCHMARK

A. Simulation Benchmark

Constructing a large-scale dataset with real objects is time-consuming and costly. To address this challenge, we use the Breaking Bad Dataset [5], which captures natural object fragmentation across diverse categories and fracture patterns. For physics simulation, we employ the SAPIEN [22] platform along with two Franka Panda grippers as robot actuators.

B. Real-World Benchmark

We construct a real-world benchmark (Fig.3) to standardize evaluation and facilitate reproducibility. Objects are placed on a turntable surrounded by 6 ArUco markers for localization, and scanned using a smartphone camera from top-down to level views. Around 300 frames are processed by COLMAP[23], [24] for camera pose estimation. We use Grounded SAM 2 [25], [26] and Depth Anything V2 [27] to generate masks and monocular depth, and use SDFStudio [28], [29] with depth ranking loss [30] to reconstruct object mesh. Our dataset includes everyday objects (e.g., wine glasses, mugs, bowls, teapots) sourced from global brands. Shapes vary in size, geometry, transparency, texture, and seam structure to ensure diversity.

IV. EXPERIMENTS

A. Simulation and Settings

We use EverydayColorPieces subset of Breaking Bad Dataset [5], with 15 categories, 445 shapes, and 11,820 fragment pairs. Shapes from 10 training categories are split into seen and novel instances, and 5 categories are held out to evaluate object- and category-level generalization.

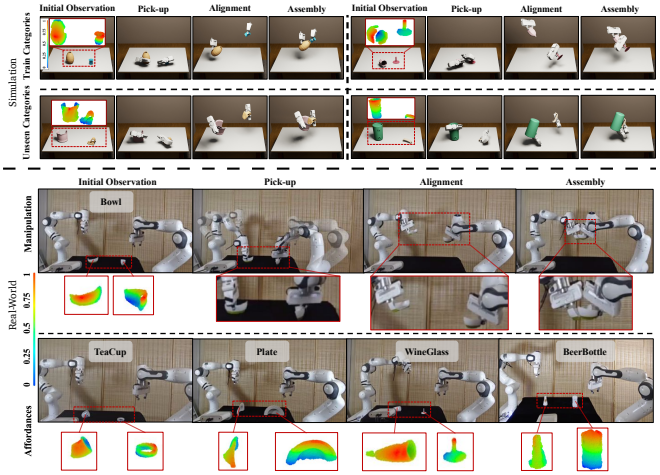

Fig. 4. **Simulation and Real-World Experiments.** We show qualitative results of the predicted affordance maps and robot actions from our method.

B. Baselines

We compare with four baselines: (1) ACT [31]: a transformer model with action chunking for closed-loop imitation of demonstrations; (2) Heuristic, a manually designed strategy that improves manipulation success; (3) SE(3)-Equiv [4]: which learns SE(3)-equivariant representations for pose estimation in vision tasks; (4) DualAfford [20]: which predicts collaborative affordance maps for bimanual manipulation.

C. Quantitative and Qualitative Results

Table I shows that our method consistently outperforms baselines across both novel instance and unseen category datasets, demonstrating strong generalization in robotic geometric assembly. Leveraging SE(3)-equivariant representations and disassembly-aware affordances, our model predicts stable grasping actions optimized for alignment and assembly. Figure 4 shows that the learned collaborative affordance maps highlight geometry-aware grasp regions while avoiding problematic areas such as fractured seams or table collisions. Our model reliably completes multi-step manipulation tasks across diverse and unseen shapes, validating its effectiveness in long-horizon bimanual assembly scenarios.

D. Real-World Experiments

As shown in Fig.4 and Fig.1 (bottom), our method performs well in real-world settings, accurately predicting grasp regions while avoiding fractured seams and areas near the table to reduce collisions.

REFERENCES

[1] G. Zhan, Q. Fan, K. Mo, L. Shao, B. Chen, L. J. Guibas, H. Dong, *et al.*, "Generative 3d part assembly via dynamic graph learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6315–6326, 2020.

[2] M. Heo, Y. Lee, D. Lee, and J. J. Lim, "Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation," *arXiv preprint arXiv:2305.12821*, 2023.

[3] Y. Lee, E. S. Hu, and J. J. Lim, "Ikea furniture assembly environment for long-horizon complex manipulation tasks," in *2021 ieee international conference on robotics and automation (icra)*. IEEE, 2021, pp. 6343–6349.

[4] R. Wu, C. Tie, Y. Du, Y. Zhao, and H. Dong, "Leveraging se (3) equivariance for learning 3d geometric shape assembly," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 14 311–14 320.

[5] S. Sellán, Y.-C. Chen, Z. Wu, A. Garg, and A. Jacobson, "Breaking bad: A dataset for geometric fracture and reassembly," in *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[6] J. Lu, Y. Sun, and Q. Huang, "Jigsaw: Learning to assemble multiple fractured objects," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[7] L. Ankile, A. Simeonov, I. Shenfeld, and P. Agrawal, "Juicer: Data-efficient imitation learning for robotic assembly," *arXiv preprint arXiv:2404.03729*, 2024.

[8] M. Yu, L. Shao, Z. Chen, T. Wu, Q. Fan, K. Mo, and H. Dong, "Roboassembly: Learning generalizable furniture assembly policy in a novel multi-robot contact-rich simulation environment," *arXiv preprint arXiv:2112.10143*, 2021.

[9] R. Wang, Y. Zhang, J. Mao, R. Zhang, C.-Y. Cheng, and J. Wu, "Ikea-manual: Seeing shape assembly step by step," *Advances in Neural Information Processing Systems*, vol. 35, pp. 28 428–28 440, 2022.

[10] F. Suárez-Ruiz, X. Zhou, and Q.-C. Pham, "Can robots assemble an ikea chair?" *Science Robotics*, vol. 3, no. 17, p. eaat6385, 2018.

[11] Z. Xian, P. Lertkultanon, and Q.-C. Pham, "Closed-chain manipulation of large objects by multi-arm robotic systems," *IEEE Robotics and Automation Letters*, vol. 2, no. 4, pp. 1832–1839, 2017.

[12] J. Lu, Y. Liang, H. Han, J. Hua, J. Jiang, X. Li, and Q. Huang, "A survey on computational solutions for reconstructing complete objects by reassembling their fractured parts," *arXiv preprint arXiv:2410.14770*, 2024.

[13] G. Papaioannou and E.-A. Karabassi, "On the automatic assemblage of arbitrary broken solid artefacts," *Image and Vision Computing*, vol. 21, no. 5, pp. 401–412, 2003.

[14] B. Liu, X. Luo, R. Huang, C. Wan, B. Zhang, W. Hu, and Z. Yue, "Virtual plate pre-bending for the long bone fracture based on axis pre-alignment," *Computerized medical imaging and graphics*, vol. 38, no. 4, pp. 233–244, 2014.

[15] J. A. Clarke, C. P. Tambussi, J. I. Noriega, G. M. Erickson, and R. A. Ketcham, "Definitive fossil evidence for the extant avian radiation in the cretaceous," *Nature*, vol. 433, no. 7023, pp. 305–308, 2005.

[16] S. Sellán, J. Luong, L. Mattos Da Silva, A. Ramakrishnan, Y. Yang, and A. Jacobson, "Breaking good: Fracture modes for realtime destruction," *ACM Transactions on Graphics*, vol. 42, no. 1, pp. 1–12, 2023.

[17] Y. Qi, Y. Ju, T. Wei, C. Chu, L. L. Wong, and H. Xu, "Two by two: Learning multi-task pairwise objects assembly for generalizable robot manipulation," *CVPR 2025*, 2025.

[18] R. Wu, Y. Zhao, K. Mo, Z. Guo, Y. Wang, T. Wu, Q. Fan, X. Chen, L. Guibas, and H. Dong, "Vat-mart: Learning visual action trajectory proposals for manipulating 3d articulated objects," *ICLR*, 2022.

[19] R. Wu, C. Ning, and H. Dong, "Learning foresightful dense visual affordance for deformable object manipulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10 947–10 956.

[20] Y. Zhao, R. Wu, Z. Chen, Y. Zhang, Q. Fan, K. Mo, and H. Dong, "Dualafford: Learning collaborative visual affordance for dual-gripper manipulation," *arXiv preprint arXiv:2207.01971*, 2022.

[21] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "Foundationpose: Unified 6d pose estimation and tracking of novel objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 868–17 879.

[22] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang, *et al.*, "Sapien: A simulated part-based interactive environment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 097–11 107.

[23] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[24] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise view selection for unstructured multi-view stereo," in *European Conference on Computer Vision (ECCV)*, 2016.

[25] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, and L. Zhang, "Grounded sam: Assembling open-world models for diverse visual tasks," 2024.

[26] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024. [Online]. Available: https://arxiv.org/abs/2408.00714

[27] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *arXiv:2406.09414*, 2024.

[28] Z. Yu, A. Chen, B. Antic, S. Peng, A. Bhattacharyya, M. Niemeyer, S. Tang, T. Sattler, and A. Geiger, "Sdfstudio: A unified framework for surface reconstruction," 2022. [Online]. Available: https://github.com/autonomousvision/sdfstudio

[29] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, "Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," *NeurIPS*, 2021.

[30] G. Wang, Z. Chen, C. C. Loy, and Z. Liu, "Sparsenerf: Distilling depth ranking for few-shot novel view synthesis," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[31] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.