BIASSEMBLE: LEARNING COLLABORATIVE AFFORDANCE FOR BIMANUAL GEOMETRIC ASSEMBLY

Anonymous authors
Paper under double-blind review

ABSTRACT

Shape assembly, the process of combining parts into a complete whole, is a crucial skill for robots with broad real-world applications. Among the 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.

1 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 (Zhan et al., 2020; Heo et al., 2023; Lee et al., 2021) and geometric assembly (Wu et al., 2023c; Sellán et al., 2022; Lu et al., 2024c). 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 (Zhan et al., 2020) to robotic systems that assemble parts in both simulation (Ankile et al., 2024; Yu et al., 2021; Wang et al., 2022a) and real-world environments (Heo et al., 2023; Suárez-Ruiz et al., 2018; Xian et al., 2017)—geometric assembly remains under-explored despite its significant potential for real-world applications (Sellán et al., 2022; Lu et al., 2024b), such as repairing broken household items, reconstructing archaeological artifacts (Papaioannou & Karabassi, 2003), assembling irregularly shaped objects in industrial tasks, aligning bone fragments in surgery (Liu et al., 2014), and reconstructing fossils in paleontology (Clarke et al., 2005).

Previous works on geometric assembly primarily focused on generating physically plausible broken parts through precise physics simulations in the graphics domain Sellán et al. (2022; 2023), and estimating the target assembled part poses based on observations in the computer vision domain Wu et al. (2023c); Lu et al. (2024c). 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.

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.

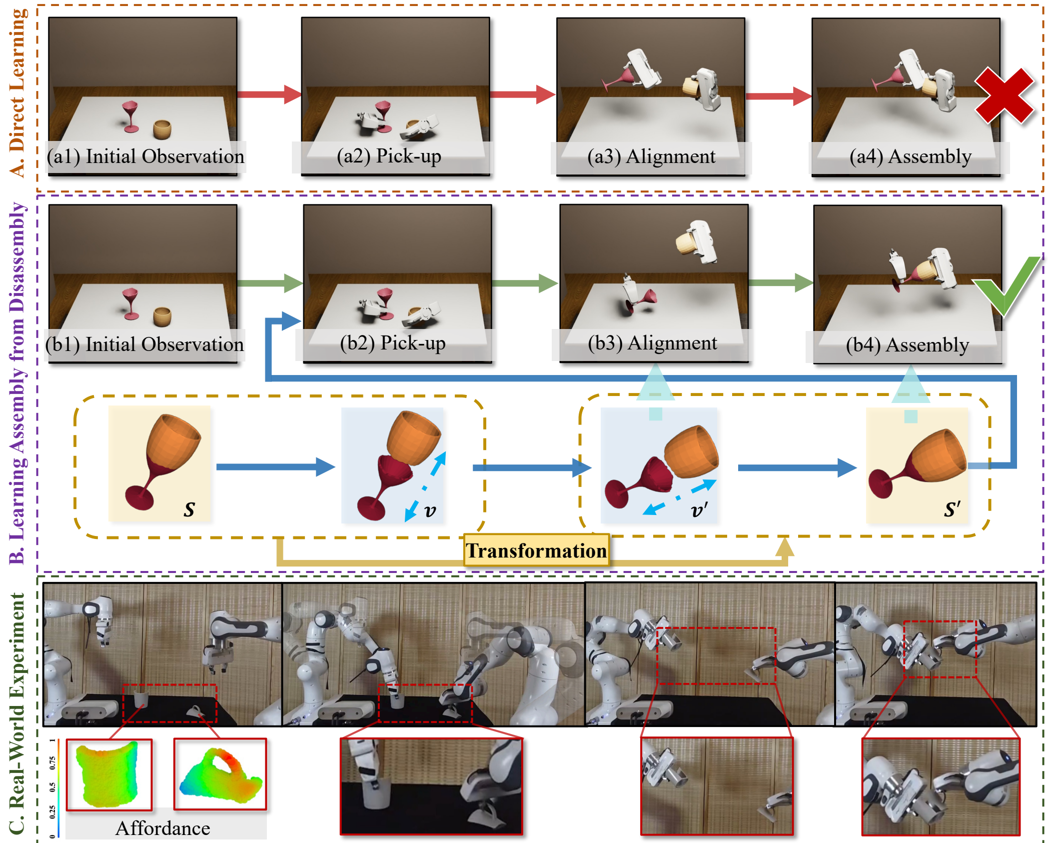

Figure 1: **(A)** Direct learning long-horizon action trajectories of geometric assembly may face many challenges: grasping ungraspable points, grasping points not suitable for assembly (*e.g.*, seams of fragments), robot colliding with parts and the other robot. **(B)** We formulate this task into 3 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.

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 Wu et al. (2022; 2023b), including short-term bimanual manipulation Zhao et al. (2022), such as pushing a box or picking up 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 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 their 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 both quantitatively and qualitatively. More results can be found in our supplementary video or on our website.

2 Related Work

2.1 3D Shape Assembly

Shape assembly is a well-established problem in visual manipulation, with many studies focusing on constructing a complete shape from given parts. These typically involve predicting the pose of each part for accurate placement using techniques like Dynamic Graph Learning (Zhan et al., 2020), or providing step-by-step guidance through human-designed visual manuals (Wang et al., 2022a). Further work (Heo et al., 2023; Tian et al., 2022; Jones et al., 2021; Willis et al., 2022) studied assembly with robotic execution, requiring robots to carry out each step. These studies offer benchmarks spanning various applications, from home furniture assembly (Lee et al., 2021) to factory-based nut-and-bolt interactions (Narang et al., 2022). We categorize these tasks into two types: furniture assembly and geometric assembly. In this paper, we focus on geometric assembly, which involves assembling pieces that are less semantically defined as individual parts. For example, in the case of a broken bowl, the pieces are irregular in shape and lack specific names, making categorization difficult. This contrasts with furniture assembly, where each piece, like a nut, bolt, or screw, has a distinct function and is named accordingly, with specific roles in the overall construction.

Previous work on geometric assembly (Sellán et al., 2022; Chen et al., 2022; Wu et al., 2023c; Lu et al., 2024c; Lee et al., 2024), such as (Wu et al., 2023c), learns SE(3)-equivariant part representations by capturing part correlations for multi-part assembly, while (Lee et al., 2024) introduces Proxy Match Transform (PMT), a low-complexity, high-order feature transform layer that refines feature pair matching. However, these methods primarily focus on synthesizing parts into a cohesive object based on pose considerations without incorporating robotic execution, which is impractical in real-world scenarios where collisions may occur if the assembly process ignores actions. To overcome this challenge, we introduce the robotic bimanual geometric assembly framework. Our approach leverages two robots to collaboratively assemble pieces, enhancing stability in real-world execution.

2.2 Bimanual Manipulation.

Bimanual manipulation (Chen et al., 2023; Grannen et al., 2023; Mu et al., 2021; Chitnis et al., 2020; Lee et al., 2015; Xie et al., 2020; Ren et al., 2024b; Liu et al., 2024; 2022; Li et al., 2023; Mu et al., 2024) offers several advantages, particularly in tasks requiring stable control or wide action space. Current research in this field primarily focuses on planning and collaboration. For instance, ACT (Fu et al., 2024; Zhao et al., 2023) introduces a transformer-based encoder-decoder architecture that leverages semantic knowledge from image inputs to predict joint positions for both arms in the next time step. PerAct2 (Grotz et al., 2024) learns features at both voxel and language levels, utilizing shared and private transformer blocks to coordinate two robotic arms based on semantic instructions, such as 'bring me a coke.' However, in tasks rich in geometric complexity, where objects have limited semantic information but intricate geometric structures, these approaches—focused on semantic understanding—may encounter generalization limits. DualAfford (Zhao et al., 2022) learns point-level collaborative visual actionable affordance, while only for short-term tasks like pushing or rotating. To address this, we leverage the geometric generalization capability of point-level affordance, and enhance it with the awareness of subsequent long-horizon assembly actions.

2.3 Visual Affordance for Robotic Manipulation

Among various vision-based approaches for robotic manipulation An et al. (2024); Goyal et al. (2023); Brohan et al. (2023); Ze et al. (2024); Ju et al. (2024), for objects with rich geometric information and tasks requiring geometric generalization, point-level affordance, which reflects the functionality of each point for downstream manipulation (Mo et al., 2021; Li et al., 2024a), is broadly leveraged and can easily generalize to novel shapes with similar local geometries. A series of research have leverage this representation to a broad range of robotic manipulation tasks, such as deformable object manipulation (Wu et al., 2023b; Lu et al., 2024a; Wu et al., 2024), object manipulation in complex environments (Ding et al., 2024; Li et al., 2024b; Wu et al., 2023a), object manipulation with efficient exploration (Ning et al., 2024; Wang et al., 2022b), and short-term bimanual manipulation Zhao et al. (2022). Leveraging the strengths of affordance representation, we design a sophisticated approach that incorporates this representation into bimanual geometric assembly task requiring long-horizon fine-grained actions, enhancing generalization and enabling more effective collaboration in addressing long-horizon geometric assembly challenges.

3 Problem Formulation

The task is to use two grippers to assemble a pair of 3D fractured parts initialized in random poses on the table. A camera situated in front of the table captures a partially scanned point cloud O. Given O, current state-of-the-art methods Scarpellini et al. (2024); Lu et al. (2024c); Chen et al. (2022) can imagine the assembled part poses and the assembled object S in any pose. Thus we assume taking imaginary assembled shape S as the input.

For the policy π, as illustrated in Figure 1, we can simplify this long-term process into 3 key steps:

- **Pick-up**: the two grippers pick up the fractured parts with actions (g_1^{pick}, g_2^{pick});
- **Alignment**: grippers carry parts to alignment poses with actions $(g_1^{align}, g_2^{align})$, positioning part seams to face each other and ensuring precise alignment for a perfect assembly;
- **Assembly**: grippers move forward to complete the assembly with actions (g_1^{asm}, g_2^{asm}).

Here, 1 and 2 denote the left and the right grippers, respectively. Each gripper action g is formulated as an $SE(3)$ matrix, representing the gripper pose in 3D space.

4 Method

4.1 Overview

Our BiAssembly framework is designed to predict collaborative affordance and gripper actions for bimanual geometric shape assembly. As illustrated in Figure 2, BiAssembly consists of several key components. 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 (4.2). Next, we design the Transformation Predictor, to transform disaasembled parts to poses where the controller can successfully manipulate the initial parts to these alignment poses (4.3). 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 (4.4). Finally, we explain training strategy and loss functions(4.6).

4.2 Disassembly Prediction Based on Fracture Geometry

The set of feasible disassembly directions (in which the disassembly and opposite assembly processes will not result in collisions) is an inherent attribute of a pair of fractured parts, determined by fracture geometries. Therefore, we predict the disassembly directions, from the object-centric perspective, on the imaginary assembled shape S in any pose. Additionally, we observe that when fractured parts rotate, the feasible disassembly directions will rotate correspondingly, maintaining SO(3) equivariance relative to part poses. This SO(3) equivariance property is advantageous for disentangling shape geometry from shape poses, as demonstrated in previous works (Wu et al., 2023c; Scarpellini et al., 2024). Therefore, we adopt VN-DGCNN (Wu et al., 2023c; Deng et al., 2021) to encode the imaginary assembled shape parts S and acquire the SO(3)-equivariant shape feature f_s.

Inputting the equivariant representation f_s, we use the Disassembly Predictor to predict the distribution of disassembly directions. Concretely, the Disassembly Predictor is implemented as a conditional variational autoencoder (cVAE) (Sohn et al., 2015), where the cVAE encoder maps the input disassembly direction v into Gauissian noise $z \in \mathbb{R}^3 2$, and the cVAE decoder reconstructs the disassembly direction v from z, with f_s as the condition.

4.3 Transformation Prediction For Alignment Pose

Given the object-centric disassembly direction resulting in no collisions in the last-step assembly, we want to predict the alignment poses, where the robot can manipulate two parts from the initial poses to the alignment poses without collisions, and then the robot can execute the assembly step. This problem can be formulated as predicting an SE(3) transformation $M \in \mathbb{R}^{4\times4}$ that is applied to the combination of the imaginary assembled shape S and the disassembly direction v. To capture this, we adopt PointNet++ (Qi et al., 2017a;b) to encode the initial point cloud observation O into the global feature f_O. We also employ a multi-layer perception (MLP) to encode disassembly direction v into the feature f_v. The transformation predictor, which is implemented as a cVAE, takes in

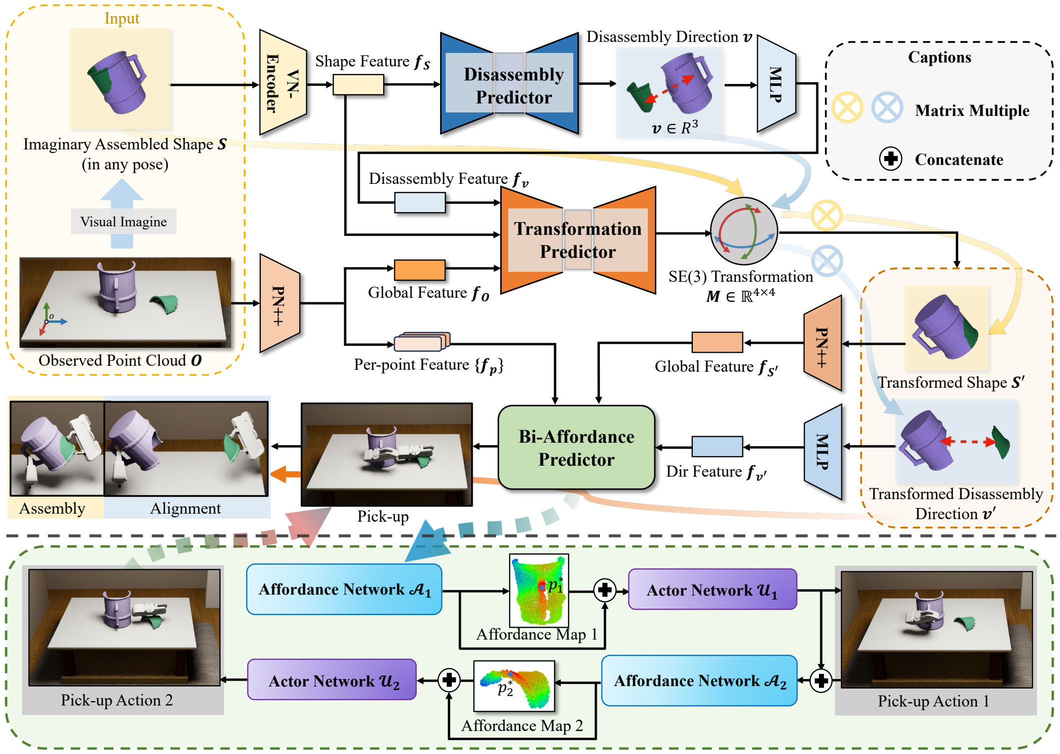

Figure 2: **BiAssembly Framework Overview.** With the point cloud observation and Imaginary Assembled Shape, the model predicts the disassembly direction in which the disassembled part poses can be easily reached by manipulating the raw parts under the guidance Bi-Affordance.

concatenation of (f_O, f_v) to predict the SE(3) transformation M. The yellow and blue arrow in Figure 2 illustrate the data flow in this process: by applying the transformation M to the imaginary assembly S and disassembly direction v, we obtain the transformed S' and v'. Therefore, we can get the target poses to which the objects should be moved during the alignment and assembly phases.

4.4 BiAffordance Predictor

We build the BiAffordance Predictor to propose actions in the **Pick-up** step, indicating where to grasp for the two fractured parts that can facilitate the whole long-horizon robotic assembly task. The BiAffordance Predictor should both identify easy-to-grasp regions on the fractured parts and consider the subsequent **Alignment** and **Assembly** steps. This means (1) avoiding grasping regions in the seam and (2) preventing each gripper from adopting poses that could collide with the other part or gripper during subsequent steps.

Following DualAfford (Zhao et al., 2022), we disentangle the bimanual task into two conditional submodules. As presented in Figure 2 (bottom), during inference, the BiAffordance Predictor conditionally predicts two gripper actions. The first Affordance Network generates the affordance map for the first gripper, highlighting the actionable regions for the bimanual assembly task, and we select a contact point p_1^* with high actionable value. Then, the Actor Network predicts the gripper orientation r_1 for interaction at p_1^*. Based on the first action $g_1 = (p_1^*, r_1)$, we can then predict the second gripper action $g_2 = (p_2^*, r_2)$ using the second Affordance Network and Actor Network.

Different from DualAfford that only predicts affordance for short-term tasks, we use whether the manipulation points can satisfy the following alignment pose (by the robotic controller) and the subsequent assembly step as the training signal.

To encode input information, one PointNet++ encodes the initial point cloud O obtains per-point features $\{f_p\}$. Another PointNet++ encodes the transformed shape S' and derives the global feature $f_{s'}$. Additionally, a MLP encodes the transformed disassembly direction v' into $f_{v'}$.

For Affordance and Actor Networks' designs, the first Affordance Network is implemented as an MLP that receives the concatenated features $(f_p, f_{S'}, f_{v'})$ and predicts an affordance score in the range of $[0, 1]$ for each point p. By aggregating the affordance scores of all points, we obtain the first affordance map, and from which we select p_1^*. The first Actor Network is implemented as a cVAE that takes the concatenated features $(f_{p_1^*}, f_{S'}, f_{v'})$ as condition, and outputs the gripper orientation r_1. The design of the second Affordance Network and Actor Network follows a similar structure, with the difference that they additionally incorporate the first gripper action's feature $(f_{p_1^*}, f_{r_1})$ along with $(f_p, f_{S'}, f_{v'})$. More details about the BiAffordance Predictor can be found in Appendix C.

4.5 Alignment and Assembly Actions

After successfully grasping a part, we now predict the gripper alignment poses g_i^{align} and assembly poses g_i^{asm}, with $i \in \{1, 2\}$ denotes the gripper id. We assume the relative pose between the gripper and the object remains stable. For example, in the first pickup step and the third assembly step, the relative gripper-object pose remains consistent, as expressed in Equation 1:

$$g_i^{pick} \cdot q_i^{pick} = g_i^{asm} \cdot q_i^{asm}; \qquad g, q \in SE(3). \tag{1}$$

Here, g and q denote the gripper and object poses, respectively.

Next, as we have the imaginary part shapes $\mathcal{P}$ with pose q_i^{init}, we can utilize a pretrained pose estimation model (Wen et al., 2024) to predict the relative pose of q_i^{pick} with respect to q_i^{init}. Besides, by applying the predicted transformation M to $\mathcal{P}$, we obtain the target assembled part $\mathcal{P}'$ and its pose as $q_i^{asm} = M \cdot q_i^{init}$. The gripper pose g_i^{pick} can be acquired from the robot control interface. Therefore, the gripper's final pose for assembling the parts can be calculated using Equation 2:

$$g_i^{asm} = g_i^{pick} \cdot q_i^{pick} \cdot \left(q_i^{init}\right)^{-1} \cdot M^{-1}; \qquad g, q \in SE(3). \tag{2}$$

It is important to note that, as indicated in the above simplified equation, we do not need to define a canonical pose or try to obtain the values of q_i^{init}; we only require the relative pose of q_i^{pick} to q_i^{init}.

A similar relationship can be established between the first and the second intermediate steps, with the difference being that $q_i^{align} = M \cdot q_i^{init} + v'$.

4.6 Training and Losses

Disassembly Direction Loss. The Disassembly Predictor is implemented as cVAE. We apply Cosine Similarity Loss to measure the error between the reconstructed disassembly direction v and ground-truth v^*, and KL Divergence to measure the difference between two distributions:

$$\mathcal{L}_{Disasm} = \mathcal{L}_{CLS}(v, v^*) + D_{KL}\big(q(z|v^*, f_s)||\mathcal{N}(0, 1)\big). \tag{3}$$

Transformation Loss. The predicted SE(3) transformation matrix M consists of translation T and rotation R. Our model predicts the translation as a 3D-vector using L1 Loss. The rotation, represented as a SO(3) matrix, can be expressed as a 6D vector by using two 3D vectors that correspond to the directions of the two orthogonal axes. Consequently, ours model predicts the rotation as a 6D-vector and employs the geodesic loss. In summary, let T^* and R^* denote the ground-truth, and for simplicity, denote $D_{KL}\big(q(z|x, f)||\mathcal{N}(0, 1)\big)$ as $D_{KL}(x, f)$. The loss function is:

$$\mathcal{L}_{Transformation} = \mathcal{L}_1(T, T^*) + \mathcal{L}_{geo}(R, R^*) + D_{KL}(T^*, (f_s, f_v)) + D_{KL}(R^*, (f_s, f_v)). \tag{4}$$

For the losses used in the Bi-Affordance Predictor, we provide detailed explanation in Appendix C.

5 Benchmark

5.1 Simulation Benchmark

Constructing a large-scale dataset with real objects is both time-consuming and costly. To address this challenge, we utilize the Breaking Bad Dataset Sellán et al. (2022), which models the natural destruction process of geometric objects into fragments. This dataset features multiple categories, diverse objects, and varying fracture patterns. For physics simulation, we employ the SAPIEN Xiang et al. (2020) platform along with two two-finger Franka Panda grippers as robot actuators.

Figure 3: **Part A** illustrates the pipeline for scanning and reconstructing real objects. **Part B** presents examples of fractured parts from various categories, showcasing diverse geometries.

We randomly select a pair of 3D fractured parts from a randomly chosen shape within a random category. The initial part poses are also randomized. Given the considerable diversity in fractured parts, collecting successful manipulation data for assembly can be quite challenging. To enhance data collection efficiency, we implement several heuristic strategies, with details in Appendix B.

5.2 Real-World Benchmark

A real-world benchmark is crucial not only for evaluating the performance of various methods but also for providing a standardized platform that enables researchers to reproduce and share their approaches. As illustrated in Figure 3, we build the real-world benchmark by scanning with a smart phone camera. First, we put the object on an automatic turntable with 6 aruco markers around for precise camera localization, and capture a RGB video from a top-down view to a level view, lowering the height by one level for each 360-degree rotation. After capturing 4-5 levels, we uniformly sample around 300 frames, and feed them to COLMAP (Schönberger & Frahm (2016); Schönberger et al. (2016)) for estimating camera poses. Then, we use Grounded SAM 2 (Ren et al. (2024a); Ravi et al. (2024)) to generate object masks and Depth Anything V2 (Yang et al. (2024)) to predict monocular depths, and use SDFStudio (Yu et al. (2022),Wang et al. (2021)) with depth ranking loss (Wang et al. (2023)) to reconstruct object mesh. To annotate the ground-truth of scanned object assembly, we import the object slices to Blender (Community (2018)) and edit the object transformations.

Our real-world benchmark encompasses a diverse range of object categories, including wine glass, plate, beer bottle, bowl, mug, and teapot. These objects have been primarily selected from well-known international brands, ensuring both durability and accessibility. To promote object diversity, our shapes vary in size, geometry, transparency, and texture, with different seam geometries.

6 Experiments

6.1 Simulation and Settings

The simulation environment is built on the SAPIEN (Xiang et al., 2020) platform, utilizing the Franka Panda grippers as the robot actuator. We employ the EverydayColorPieces dataset from the Breaking Bad Dataset Sellán et al. (2022), covering 15 categories, 445 shapes and 11,820 fragment pairs, with 10 categories for training and the remaining 5 for testing. Training categories are further divided into training shapes and novel instances, allowing the evaluation on generalization capabilities at both the object and category levels. More details can be found in Appendix A.

For each method, we provide 7,000 positive and 7,000 negative samples. The negative samples encompass manipulation failures occurring during the grasping, alignment, and assembly steps.

Table 1: Quantitative results in novel instances within training categories.

Method	Novel Instances in Training Categories										AVG
ACT	2%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0.30%
Heuristic	5%	8%	0%	3%	2%	4%	3%	5%	10%	2%	4.20%
DualAfford	21%	17%	0%	2%	2%	4%	14%	8%	10%	6%	8.40%
w/o SE(3)	59%	29%	**13%**	**14%**	11%	8%	15%	**19%**	24%	20%	21.20%
Ours	**60%**	**38%**	**13%**	13%	**12%**	**9%**	**26%**	18%	**27%**	**25%**	**24.10%**
w/ GT Target	71%	28%	4%	9%	9%	13%	27%	19%	25%	19%	22.40%

Table 2: Quantitative results in the novel unseen categories.

Method	Unseen Categories					AVG
ACT	0%	1%	0%	0%	1%	0.4%
Heuristic	1%	5%	2%	0%	14%	4.4%
DualAfford	5%	10%	4%	1%	16%	7.2%
w/o SE(3)	13%	24%	4%	**9%**	22%	14.4%
Ours	**14%**	**31%**	**10%**	7%	**25%**	**17.4%**
w/ GT Target	14%	33%	12%	9%	27%	19%

6.2 Evaluation Metric, Baselines and Ablation

Evaluation Metric. Our metric evaluates whether the relative distance (measured in unit-length) and rotation angle (measured in degrees) of two parts are within the threshold range at the end of the assembly. These thresholds ensure that the success of the assembly process can be measured consistently and meaningfully. To evaluate each method, we prepare 100 samples in each category. For each sample, all methods are presented with the same initial observation for a fair comparison.

Baselines and Ablations. We compare our approach with three baselines and two ablated version: (1) ACT (Zhao et al., 2023), a transformer network with action chunking that imitates successful action sequences in the closed-loop manner. We enhance this method by providing depth information, object pose, and an additional target goal image as inputs. Besides, this method is trained and tested on individual categories, whereas other learning-based methods are trained on all training categories and evaluated on both the novel instances and unseen categories. (2) Heuristic, where we hand-engineer a set of heuristic strategies to improve manipulation success rate. These strategies are similar to the data collection heuristics described in Appendix B. (3) DualAfford Zhao et al. (2022), a framework that learns collaborative visual affordance for bimanual manipulation. While DualAfford focuses on short-term manipulation, we adapt it to determine where to grasp the two fractured parts, using heuristic methods for the alignment and assembly steps. (4) w/o SE(3), an ablated version that replaces the SO(3)-equivariant VN-DGCNN encoder with PointNet++. (5) w/ GT target, where we provide the additional ground-truth disassembly direction v and transformation M. The ground truth is sampled using a heuristic method that ensures at least one feasible assembly.

6.3 Quantitative Results and Analysis

Table 1 and Table 2 show the success rate comparisons across different methods on both the novel instance dataset and the unseen category dataset. Our method outperforms the baselines and ablation models in most cases, demonstrating its effectiveness and geometric generalization capabilities. For **ACT**, though we provide additional input such as depth, object poses, and the goal image, it achieves lower scores on our task. Although ACT successfully picks up parts in approximately 40% of trials, it often fails during the alignment phase, with a misalignment of over 100° between the parts in many cases. Furthermore, ACT struggles to avoid grasping the fractured seam regions, leading to collisions during the assembly process. This is because the observation and action spaces in robotic geometric assembly are exceptionally large, making it challenging to directly learn the appropriate fine-grained actions. For **heuristic**, it achieves higher scores because we provide substantial

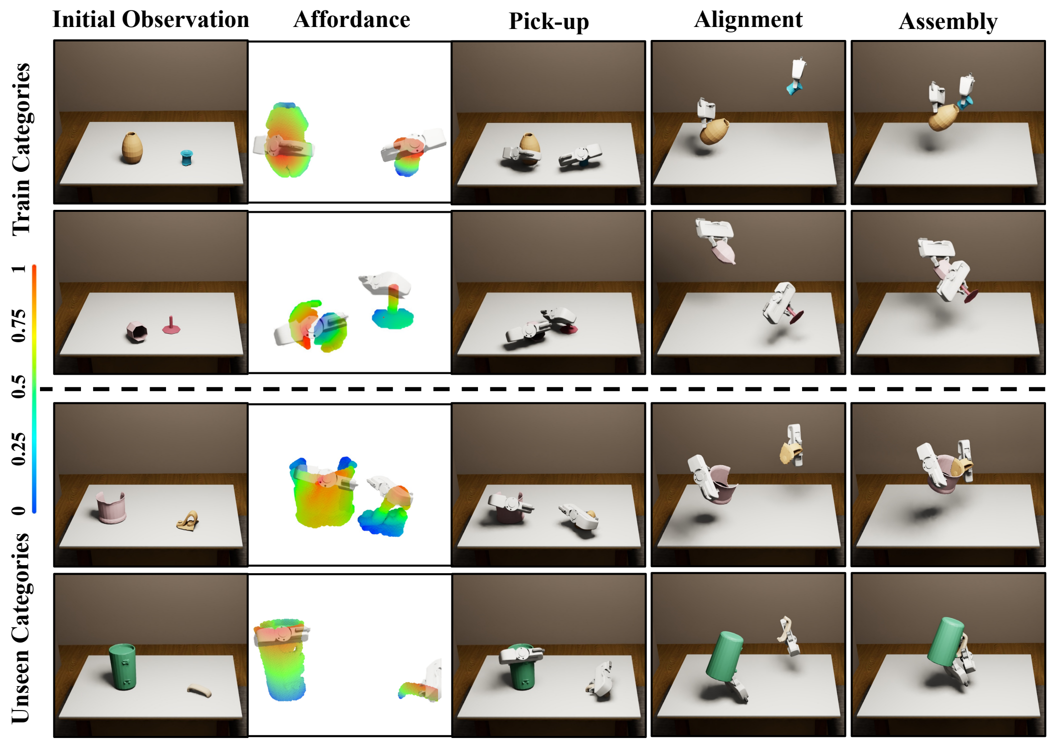

Figure 4: We present qualitative results of the predicted affordance maps and robot actions from our method. In each row, from left to right, we respectively present the input observation, the predicted affordance maps for the two fractured parts, and the bimanual actions for the pick-up, alignment, and assembly steps. In the top part are novel shapes from the training categories, while in the bottom part are shapes from unseen categories.

ground-truth information in the simulation. However, due to the significant diversity in both inter- and intra-category shape geometries, it is unrealistic to expect hand-engineered rules to generalize effectively across all shapes. **DualAfford** performs better than the heuristic policy, demonstrating its superior ability to learn geometric-aware pick-up poses compared to heuristic sampling. However, only with designs focused on short-term manipulation, it still lacks awareness of subsequent alignment and assembly steps. Comparing our method to the ablation **w/o SE(3)**, we observe that utilizing the SE(3)-equivariant representation enhances the performance across most categories. Lastly, compared to the ablation **w/ GT Target**, our method performs better on novel instances but worse on novel categories. This suggests that, for the training categories, our method learns to predict a more accurate distribution of disassembly and transformation, surpassing those sampled from the heuristic strategy. However, on novel unseen categories, while our method still demonstrates generalization capability, the ablated version with the ground-truth target remains more effective.

6.4 Qualitative Results and Analysis

In Figure 4, we present the collaborative affordance maps and robot manipulations predicted by our methods across multiple categories, including novel instances in the training categories and the unseen categories. The predicted affordance demonstrates an awareness of part geometry, highlighting graspable regions while avoiding areas near the table that could result in collisions between the gripper and the surface. Additionally, the affordance accounts for subsequent alignment and assembly steps, avoiding seam areas that may cause collisions during the approach phase. Based on the predicted affordance map, our model predicts appropriate gripper actions for assembling parts. Moreover, the results demonstrate the model's ability to generalize to unseen categories and shapes.

6.5 Real-World Experiments

We set up two Franka Panda robots with the fractured parts positioned between them. An Azure Kinect camera is mounted in front of the robots, capturing partial 3D point cloud data as inputs

Figure 5: **Real-World Experiment.** We present the results of our model tested on real-world scans. For each data, We visualized the affordance map, and the bimanual actions for the pick-up, alignment, and assembly steps. Manipulation videos can be found in our supplementary materials.

for our models. The robots are controlled via the Robot Operating System (ROS) (Quigley et al., 2009), with control and communication managed through the frankapy library (Zhang et al., 2020). Communication with the Kinect Azure is facilitated by the pyk4a library (pyk4a, 2019).

In the bottom row of Figure 1 and in Figure 5, we present promising results by directly testing our method in real-world scenarios. We observe that our model not only learns which regions of the fractured parts to grasp but also avoids manipulating areas near the fracture regions or too close to the table surface, reducing the likelihood of collisions during manipulation. The results from the real-world experiments demonstrate our model's capacity for generalization to real-world scenarios.

7 Conclusion

In conclusion, we have leveraged the geometric generalization capability of point-level affordance to develop a method that enables both generalization and collaboration in long-horizon geometric assembly tasks. To evaluate performance across diverse geometries, we introduced a real-world benchmark that features significant geometric variety and global reproducibility. Extensive experiments have shown that our approach outperforms previous methods, demonstrating its effectiveness in handling complex and long-horizon assembly tasks. For more discussions, including potential extensions to multi-part shape assembly and future directions, are detailed in Appendix F.

REFERENCES

Boshi An, Yiran Geng, Kai Chen, Xiaoqi Li, Qi Dou, and Hao Dong. Rgbmanip: Monocular image-based robotic manipulation through active object pose estimation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7748–7755. IEEE, 2024.

Lars Ankile, Anthony Simeonov, Idan Shenfeld, and Pulkit Agrawal. Juicer: Data-efficient imitation learning for robotic assembly. *arXiv preprint arXiv:2404.03729*, 2024.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Yuanpei Chen, Yiran Geng, Fangwei Zhong, Jiaming Ji, Jiechuang Jiang, Zongqing Lu, Hao Dong, and Yaodong Yang. Bi-dexhands: Towards human-level bimanual dexterous manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Yun-Chun Chen, Haoda Li, Dylan Turpin, Alec Jacobson, and Animesh Garg. Neural shape mating: Self-supervised object assembly with adversarial shape priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12724–12733, 2022.

Rohan Chitnis, Shubham Tulsiani, Saurabh Gupta, and Abhinav Gupta. Efficient bimanual manipulation using learned task schemas. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1149–1155. IEEE, 2020.

Julia A Clarke, Claudia P Tambussi, Jorge I Noriega, Gregory M Erickson, and Richard A Ketcham. Definitive fossil evidence for the extant avian radiation in the cretaceous. *Nature*, 433(7023): 305–308, 2005.

Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. URL `http://www.blender.org`.

Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas J Guibas. Vector neurons: A general framework for so (3)-equivariant networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12200–12209, 2021.

Kairui Ding, Boyuan Chen, Ruihai Wu, Yuyang Li, Zongzheng Zhang, Huan-ang Gao, Siqi Li, Yixin Zhu, Guyue Zhou, Hao Dong, et al. Preafford: Universal affordance-based pre-grasping for diverse objects and environments. *IROS*, 2024.

Zipeng Fu, Tony Z Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *arXiv preprint arXiv:2401.02117*, 2024.

Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. Rvt: Robotic view transformer for 3d object manipulation. In *Conference on Robot Learning*, pp. 694–710. PMLR, 2023.

Jennifer Grannen, Yilin Wu, Brandon Vu, and Dorsa Sadigh. Stabilize to act: Learning to coordinate for bimanual manipulation. In *Conference on Robot Learning*, pp. 563–576. PMLR, 2023.

Markus Grotz, Mohit Shridhar, Tamim Asfour, and Dieter Fox. Peract2: A perceiver actor framework for bimanual manipulation tasks. *arXiv preprint arXiv:2407.00278*, 2024.

Minho Heo, Youngwoon Lee, Doohyun Lee, and Joseph J Lim. Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation. *arXiv preprint arXiv:2305.12821*, 2023.

Benjamin Jones, Dalton Hildreth, Duowen Chen, Ilya Baran, Vladimir G Kim, and Adriana Schulz. Automate: A dataset and learning approach for automatic mating of cad assemblies. *ACM Transactions on Graphics (TOG)*, 40(6):1–18, 2021.

Yuanchen Ju, Kaizhe Hu, Guowei Zhang, Gu Zhang, Mingrun Jiang, and Huazhe Xu. Robo-abc: Affordance generalization beyond categories via semantic correspondence for robot manipulation. *arXiv preprint arXiv:2401.07487*, 2024.

Alex X Lee, Henry Lu, Abhishek Gupta, Sergey Levine, and Pieter Abbeel. Learning force-based manipulation of deformable objects from multiple demonstrations. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 177–184. IEEE, 2015.

Nahyuk Lee, Juhong Min, Junha Lee, Seungwook Kim, Kanghee Lee, Jaesik Park, and Minsu Cho. 3d geometric shape assembly via efficient point cloud matching. *arXiv preprint arXiv:2407.10542*, 2024.

Youngwoon Lee, Edward S Hu, and Joseph J Lim. Ikea furniture assembly environment for long-horizon complex manipulation tasks. In *2021 ieee international conference on robotics and automation (icra)*, pp. 6343–6349. IEEE, 2021.

Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18061–18070, 2024a.

Yitong Li, Ruihai Wu, Haoran Lu, Chuanruo Ning, Yan Shen, Guanqi Zhan, and Hao Dong. Broadcasting support relations recursively from local dynamics for object retrieval in clutters. *RSS*, 2024b.

Yunfei Li, Chaoyi Pan, Huazhe Xu, Xiaolong Wang, and Yi Wu. Efficient bimanual handover and rearrangement via symmetry-aware actor-critic learning. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3867–3874. IEEE, 2023.

Bin Liu, Xinjian Luo, Rui Huang, Chao Wan, Bingbing Zhang, Weihua Hu, and Zongge Yue. Virtual plate pre-bending for the long bone fracture based on axis pre-alignment. *Computerized medical imaging and graphics*, 38(4):233–244, 2014.

Junjia Liu, Yiting Chen, Zhipeng Dong, Shixiong Wang, Sylvain Calinon, Miao Li, and Fei Chen. Robot cooking with stir-fry: Bimanual non-prehensile manipulation of semi-fluid objects. *IEEE Robotics and Automation Letters*, 7(2):5159–5166, 2022.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21740–21751, 2024.

Haoran Lu, Yitong Li, Ruihai Wu, Chuanruo Ning, Yan Shen, and Hao Dong. Unigarment: A unified simulation and benchmark for garment manipulation. In *ICRA Workshop on Deformable Object Manipulation*, 2024a.

Jiaxin Lu, Yongqing Liang, Huijun Han, Jiacheng Hua, Junfeng Jiang, Xin Li, and Qixing Huang. A survey on computational solutions for reconstructing complete objects by reassembling their fractured parts. *arXiv preprint arXiv:2410.14770*, 2024b.

Jiaxin Lu, Yifan Sun, and Qixing Huang. Jigsaw: Learning to assemble multiple fractured objects. *Advances in Neural Information Processing Systems*, 36, 2024c.

Kaichun Mo, Leonidas J Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6813–6823, 2021.

Tongzhou Mu, Zhan Ling, Fanbo Xiang, Derek Yang, Xuanlin Li, Stone Tao, Zhiao Huang, Zhiwei Jia, and Hao Su. Maniskill: Learning-from-demonstrations benchmark for generalizable manipulation skills. *CoRR, abs/2107.14483, 2021b. URL https://arxiv. org/abs/2107*, 14483, 2021.

Yao Mu, Tianxing Chen, Shijia Peng, Zanxin Chen, Zeyu Gao, Yude Zou, Lunkai Lin, Zhiqiang Xie, and Ping Luo. Robotwin: Dual-arm robot benchmark with generative digital twins (early version). *arXiv preprint arXiv:2409.02920*, 2024.

Yashraj Narang, Kier Storey, Iretiayo Akinola, Miles Macklin, Philipp Reist, Lukasz Wawrzyniak, Yunrong Guo, Adam Moravanszky, Gavriel State, Michelle Lu, et al. Factory: Fast contact for robotic assembly. *arXiv preprint arXiv:2205.03532*, 2022.

Chuanruo Ning, Ruihai Wu, Haoran Lu, Kaichun Mo, and Hao Dong. Where2explore: Few-shot affordance learning for unseen novel categories of articulated objects. *Advances in Neural Information Processing Systems*, 36, 2024.

Georgios Papaioannou and Evaggelia-Aggeliki Karabassi. On the automatic assemblage of arbitrary broken solid artefacts. *Image and Vision Computing*, 21(5):401–412, 2003.

pyk4a. pyk4a, 2019. URL `https://github.com/etiennedub/pyk4a`.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017a.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017b.

Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, Andrew Y Ng, et al. Ros: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, pp. 5. Kobe, Japan, 2009.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL `https://arxiv.org/abs/2408.00714`.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024a.

Yi Ren, Zhehua Zhou, Ziwei Xu, Yang Yang, Guangyao Zhai, Marion Leibold, Fenglei Ni, Zhengyou Zhang, Martin Buss, and Yu Zheng. Enabling versatility and dexterity of the dual-arm manipulators: A general framework toward universal cooperative manipulation. *IEEE Transactions on Robotics*, 2024b.

Gianluca Scarpellini, Stefano Fiorini, Francesco Giuliari, Pietro Morerio, and Alessio Del Bue. Diffassemble: A unified graph-diffusion model for 2d and 3d reassembly. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024.

Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

Silvia Sellán, Yun-Chun Chen, Ziyi Wu, Animesh Garg, and Alec Jacobson. Breaking bad: A dataset for geometric fracture and reassembly. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Silvia Sellán, Jack Luong, Leticia Mattos Da Silva, Aravind Ramakrishnan, Yuchuan Yang, and Alec Jacobson. Breaking good: Fracture modes for realtime destruction. *ACM Transactions on Graphics*, 42(1):1–12, 2023.

Issei Sera, Natsuki Yamanobe, Ixchel G Ramirez-Alpizar, Zhenting Wang, Weiwei Wan, and Kensuke Harada. Assembly planning by recognizing a graphical instruction manual. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3138–3145. IEEE, 2021.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28: 3483–3491, 2015.

Francisco Suárez-Ruiz, Xian Zhou, and Quang-Cuong Pham. Can robots assemble an ikea chair? *Science Robotics*, 3(17):eaat6385, 2018.

Yunsheng Tian, Jie Xu, Yichen Li, Jieliang Luo, Shinjiro Sueda, Hui Li, Karl DD Willis, and Wojciech Matusik. Assemble them all: Physics-based planning for generalizable assembly by disassembly. *ACM Transactions on Graphics (TOG)*, 41(6):1–11, 2022.

Theodore Tsesmelis, Luca Palmieri, Marina Khoroshiltseva, Adeela Islam, Gur Elkin, Ofir Itzhak Shahar, Gianluca Scarpellini, Stefano Fiorini, Yaniv Ohayon, Nadav Alali, et al. Re-assembling the past: The repair dataset and benchmark for real world 2d and 3d puzzle solving. *arXiv preprint arXiv:2410.24010*, 2024.

Yuxuan Wan, Kaichen Zhou, Hao Dong, et al. Scanet: Correcting lego assembly errors with self-correct assembly network. *arXiv preprint arXiv:2403.18195*, 2024.

Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *NeurIPS*, 2021.

Ruocheng Wang, Yunzhi Zhang, Jiayuan Mao, Ran Zhang, Chin-Yi Cheng, and Jiajun Wu. Ikea-manual: Seeing shape assembly step by step. *Advances in Neural Information Processing Systems*, 35:28428–28440, 2022a.

Yian Wang, Ruihai Wu, Kaichun Mo, Jiaqi Ke, Qingnan Fan, Leonidas J Guibas, and Hao Dong. Adaafford: Learning to adapt manipulation affordance for 3d articulated objects via few-shot interactions. In *European conference on computer vision*, pp. 90–107. Springer, 2022b.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17868–17879, 2024.

Karl DD Willis, Pradeep Kumar Jayaraman, Hang Chu, Yunsheng Tian, Yifei Li, Daniele Grandi, Aditya Sanghi, Linh Tran, Joseph G Lambourne, Armando Solar-Lezama, et al. Joinable: Learning bottom-up assembly of parametric cad joints. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15849–15860, 2022.

Ruihai Wu, Yan Zhao, Kaichun Mo, Zizheng Guo, Yian Wang, Tianhao Wu, Qingnan Fan, Xuelin Chen, Leonidas Guibas, and Hao Dong. Vat-mart: Learning visual action trajectory proposals for manipulating 3d articulated objects. *ICLR*, 2022.

Ruihai Wu, Kai Cheng, Yan Zhao, Chuanruo Ning, Guanqi Zhan, and Hao Dong. Learning environment-aware affordance for 3d articulated object manipulation under occlusions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. URL `https://openreview.net/forum?id=Re2NHYoZ5l`.

Ruihai Wu, Chuanruo Ning, and Hao Dong. Learning foresightful dense visual affordance for deformable object manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10947–10956, 2023b.

Ruihai Wu, Chenrui Tie, Yushi Du, Yan Zhao, and Hao Dong. Leveraging se (3) equivariance for learning 3d geometric shape assembly. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14311–14320, 2023c.

Ruihai Wu, Haoran Lu, Yiyan Wang, Yubo Wang, and Hao Dong. Unigarmentmanip: A unified framework for category-level garment manipulation via dense visual correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16340–16350, 2024.

Zhou Xian, Puttichai Lertkultanon, and Quang-Cuong Pham. Closed-chain manipulation of large objects by multi-arm robotic systems. *IEEE Robotics and Automation Letters*, 2(4):1832–1839, 2017.

Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11097–11107, 2020.

Fan Xie, Alexander Chowdhury, M De Paolis Kaluza, Linfeng Zhao, Lawson Wong, and Rose Yu. Deep imitation learning for bimanual robotic manipulation. *Advances in neural information processing systems*, 33:2327–2337, 2020.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024.

Mingxin Yu, Lin Shao, Zhehuan Chen, Tianhao Wu, Qingnan Fan, Kaichun Mo, and Hao Dong. Roboassembly: Learning generalizable furniture assembly policy in a novel multi-robot contact-rich simulation environment. *arXiv preprint arXiv:2112.10143*, 2021.

Zehao Yu, Anpei Chen, Bozidar Antic, Songyou Peng, Apratim Bhattacharyya, Michael Niemeyer, Siyu Tang, Torsten Sattler, and Andreas Geiger. Sdfstudio: A unified framework for surface reconstruction, 2022. URL `https://github.com/autonomousvision/sdfstudio`.

Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy. *arXiv preprint arXiv:2403.03954*, 2024.

Guanqi Zhan, Qingnan Fan, Kaichun Mo, Lin Shao, Baoquan Chen, Leonidas J Guibas, Hao Dong, et al. Generative 3d part assembly via dynamic graph learning. *Advances in Neural Information Processing Systems*, 33:6315–6326, 2020.

Kevin Zhang, Mohit Sharma, Jacky Liang, and Oliver Kroemer. A modular robotic arm control stack for research: Franka-interface and frankapy. *arXiv preprint arXiv:2011.02398*, 2020.

Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

Yan Zhao, Ruihai Wu, Zhehuan Chen, Yourong Zhang, Qingnan Fan, Kaichun Mo, and Hao Dong. Dualafford: Learning collaborative visual affordance for dual-gripper manipulation. *arXiv preprint arXiv:2207.01971*, 2022.

Appendix

A Details about Data Statistics

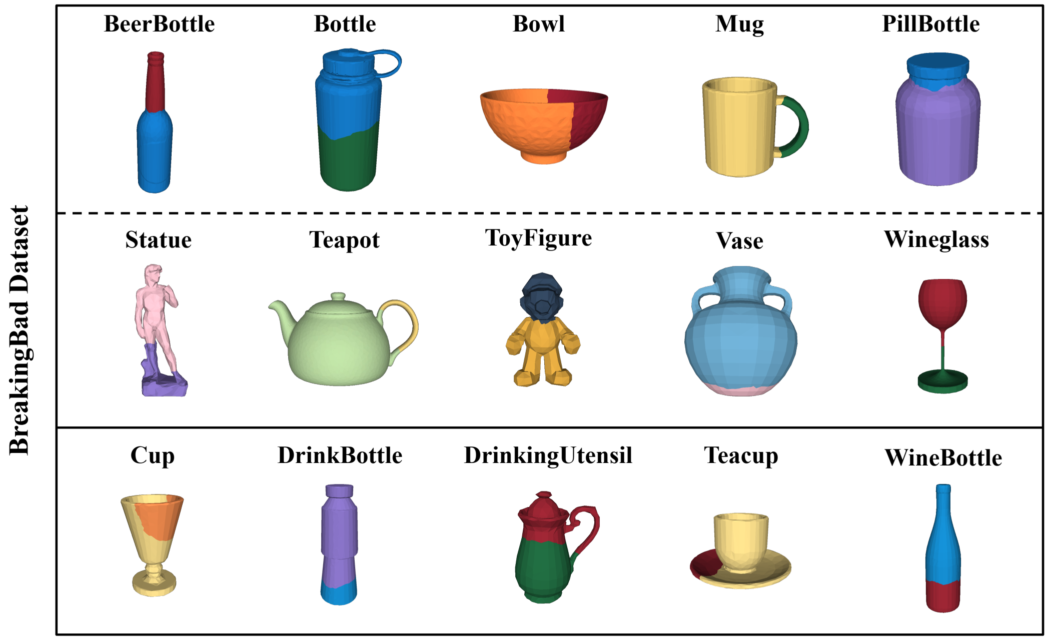

Figure 6: Visualization of simulation data. We present one example shape from each object category used in our paper.

Table 3: Shape and Fracture Counts Across Categories. Numbers before the slash represent the training set, and numbers after the slash represent the testing set. The top 10 categories are the training categories, and the bottom 5 categories are the unseen categories.

Category	Shape (Train/Test)	Fracture (Train/Test)
BeerBottle	6 / 3	100 / 61
Bottle	51 / 22	1296 / 559
Bowl	16 / 32	446 / 801
Mug	32 / 15	876 / 545
PillBottle	7 / 3	217 / 60
Statue	2 / 0	57 / 35
Teapot	7 / 3	315 / 104
ToyFigure	36 / 16	1118 / 556
Vase	74 / 32	1842 / 872
WineGlass	6 / 3	136 / 45
Cup	0 / 31	0 / 663
DrinkBottle	0 / 7	0 / 230
DrinkingUtensil	0 / 14	0 / 343
Teacup	0 / 7	0 / 167
WineBottle	0 / 18	0 / 376
Total	237 / 208	6403 / 5417

In Figure 6, we present a representative example for each object category from the dataset used in our experiments.

In this paragraph, we detail the data split for our experiments. We randomly select 10 out of the 15 categories for training, reserving the remaining 5 categories exclusively for testing. Within the 10 training categories, 60% of the shapes are randomly chosen for the training set, while the remaining 40% serve as a test set to assess the models' performance on novel instances within the training categories (shape-level). For the reserved 5 categories, all shapes are included in the test set to evaluate the methods' generalization capabilities on unseen categories (category-level). In summary, the training set consists of 10 categories, totaling 237 shapes and 6,403 pairs of fragments. The shape-level test set includes 10 categories, comprising 131 shapes and 3,638 pairs of fragments. The category-level test set encompasses 5 categories, containing 77 shapes and 1,779 pairs of fragments. Detailed statistics for each category can be found in Table 3.

B Details about Data Collection in Simulation

In this section, we provide detailed information about data collection in the simulation.

Due to the complexity of bimanual geometric assembly tasks, which stems from the vast observation and action spaces, it is nearly impossible to directly acquire positive data by randomly manipulating the fractured parts. To address this, we apply several heuristic strategies to improve the efficiency of data collection. Specifically, our strategies focus on the following three key steps in the process:

1. Sampling the grasping poses for the two grippers
2. Sampling the alignment poses for the two grippers
3. Sampling the assembly poses for the two grippers

Each of these steps is described in detail in the following subsections.

B.1 Sampling Grasping Poses

Different from furniture assembly task, where grasp points are easier to define, the objects in geometric shape assembly tasks have more diverse geometries, making it challenging to establish a consistent grasping policy. As a result, our heuristic strategy for grasping primarily focuses on the orientation of the grippers rather than specific grasp points.

At initialization, the two parts are randomly placed on the table. From the simulation, we obtain the ground-truth depth map and normalization map of the two parts. The normal directions often closely align with feasible grasping directions (i.e., the z-axis of the gripper). Consequently, we randomly select a grasp point on the part, and then choose a grasping direction within a cone that deviates less than 30 degrees from the normal direction at that point. To avoid potential collisions between the grippers and the table during grasping, we discard any directions that point towards the upper hemisphere of the world coordinate system.

In addition to the gripper's z-axis, the x-axis also significantly impacts grasping accuracy. Therefore, we uniformly sample a list of n x-axis candidates that are orthogonal to the gripper's z-axis. By combining each candidate x-axis with the z-axis, we determine the gripper pose. We test each of these gripper poses sequentially. If a grasp pose successfully grasps the object, we proceed to the next stage; otherwise, we reset the scene and move on to the next x-axis candidate. If all grasp pose candidates fail, we record this as negative data for the grasping step. In our implementation, we empirically set $n = 6$, resulting in each x-axis candidate being spaced 60 degrees apart.

B.2 Sampling Alignment Poses

To sample the grippers' alignment poses, we begin by sampling feasible part poses during the alignment step. Our heuristic strategy also follows a reverse disassembly process. Specifically, we load the ground-truth assembled object into the simulation at a height of 0.5 meters above the tabletop, and allow the assembled object to take any pose rather than being restricted to a canonical pose. Next, we randomly explore feasible disassembly directions for the two parts, ensuring that these directions are collision-free. The resulting poses of the parts, after moving in their respective disassembly directions, represent the parts' alignment poses. It is important to note that we will discard alignment poses that are too distant from the parts' initial poses (for example, if the initial left part has an alignment pose to the right, while the initial right part has an alignment pose to the left).

Once we have determined the parts' alignment poses, we can calculate the grippers' alignment poses using the functions described in Section 4.5.

B.3 Sampling Assembly Directions

In the previous step, we identified the feasible disassembly directions for the ground-truth assembled parts. Consequently, we can obtain the assembly directions by simply inverting these disassembly directions, allowing the two grippers to assemble the parts accordingly. However, it is important to note that this assembly process may lead to failures. This is because, although the parts can be successfully aligned in an idealized scenario without grippers, the presence of grippers increases the risk of collisions. For instance, if one gripper is positioned too close to the seam area of a fractured part, it may collide with another part or the other gripper during the assembly process.

C More Details about the BiAffordance Framework

In this section, we provide more details about the BiAffordance Predictor. Following DualAfford (Zhao et al., 2022), we decompose the bimanual cooperation task into two separate yet closely interconnected submodules, $\mathcal{M}_1$ and $\mathcal{M}_2$, which conditionally predict the first and second gripper actions, respectively.

During inference, the first module $\mathcal{M}_1$ predicts the first gripper action $g_1 = (p_1^*, r_1)$, followed by the second module $\mathcal{M}_2$, which predicts the second action $g_2 = (p_2^*, r_2)$ conditioned on g_1, as described in Section 4.4 of the main paper.

During training, $\mathcal{M}_2$ still takes the first gripper action g_1 as input, and then generates a complementary second action g_2. However, since $\mathcal{M}_1$ lacks knowledge of how g_2 will be predicted, it faces challenges in predicting a collaborative action g_1. To address this issue, we aim to make $\mathcal{M}_1$ aware of the types of actions that can be easily collaborated on. We assess the quality of $\mathcal{M}_1$'s actions by evaluating whether $\mathcal{M}_2$ can generate cooperative actions, which encourages $\mathcal{M}_1$ to predict actions with high collaborative quality. Following this approach, $\mathcal{M}_2$ guides the training of $\mathcal{M}_1$. Thus, we first train $\mathcal{M}_2$ and then use the trained $\mathcal{M}_2$ to train $\mathcal{M}_1$, ensuring cooperative predictions.

During training, each submodule $\mathcal{M}_i$ consists of three components: (1) an Affordance Network $\mathcal{A}_i$, which predicts an affordance map to indicate where interaction should occur; (2) an Actor Network $\mathcal{U}_i$, which predicts manipulation orientations to determine how to interact at the selected point; and (3) a Critic Network $\mathcal{C}_i$, which assesses the likelihood of the action's success.

To explain the training process, we begin with the more straightforward second module, $\mathcal{M}_2$, which is also the first to be trained.

The Actor Network $\mathcal{U}_2$ in $\mathcal{M}_2$ is implemented as a conditional Variational Autoencoder (cVAE). As detailed in Section 4.4, it takes concatenated input features $f_{in} = (f_p, f_{S'}, f_{v'})$ and the ground-truth feature of the first action $f_{g_1} = (f_{p_1^*}, f_{r_1})$ from the collected data. We apply a geodesic distance loss to measure the error between the reconstructed gripper orientation r_2 and the ground-truth orientation $\hat{r}_2$, along with KL divergence to quantify the difference between the two distributions:

$$\mathcal{L}_{\mathcal{U}_2} = \mathcal{L}_{geo}(r_2, \hat{r}_2) + D_{KL}\big(q(z|\hat{r}_2, f_{in}, f_{g_1})||\mathcal{N}(0,1)\big). \quad (5)$$

The Critic Network $\mathcal{C}_2$ in $\mathcal{M}_2$ is implemented as a multilayer perceptron (MLP) and evaluates how well the predicted second gripper action $g_2 = (p_2^*, r_2)$ collaborates with the first action g_1. Using the collected data along with the corresponding ground-truth interaction results r (where $r = 1$ indicates a positive interaction and $r = 0$ indicates a negative one), we train $\mathcal{C}_2$ with the standard binary cross-entropy loss:

$$\mathcal{L}_{\mathcal{C}_2} = r_j \log\big(\mathcal{C}_2(f_{in}, f_{g_1}, f_{p_2^*}, f_{r_2})\big) + (1 - r_j)\log\big(1 - \mathcal{C}_2(f_{in}, f_{g_1}, f_{p_2^*}, f_{r_2})\big). \quad (6)$$

The Affordance Network $\mathcal{A}_2$ in $\mathcal{M}_2$ is implemented as a multilayer perceptron (MLP). The predicted affordance score represents the expected success rate for executing action proposals generated by the Actor Network, which can be directly evaluated by the Critic Network. To obtain the ground-truth affordance score $\hat{a}_{p_i}$ on p_i, we use the Actor Network $\mathcal{U}_2$ to sample n gripper orientations at the point p_i and calculate the average action scores assigned by the Critic Network $\mathcal{C}_2$. We apply L1 loss to measure the error between the predicted and ground-truth affordance scores at a specific point p_i:

$$\hat{a}_{p_i} = \frac{1}{n}\sum_{j=1}^{n} \mathcal{C}_2\big(f_{in}, f_{g_1}, f_{p_i}, \mathcal{U}_2(f_{in}, f_{g_1}, f_{p_i}, z_j)\big); \qquad \mathcal{L}_{\mathcal{A}_2} = \left|\mathcal{A}_2(f_{in}, f_{g_1}, f_{p_i}) - \hat{a}_{p_i}\right|. \tag{7}$$

After training the expert model $\mathcal{M}_2 = (\mathcal{A}_2, \mathcal{U}_2, \mathcal{C}_2)$, we can utilize it to generate collaborative actions g_2 for a given g_1 predicted by $\mathcal{M}_1$. Thus, we can assess the quality of $\mathcal{M}_1$'s actions by evaluating whether $\mathcal{M}_2$ can generate cooperative actions. Specifically, to evaluate the predicted g_1, we use the trained $\mathcal{A}_2$ and $\mathcal{U}_2$ to generate multiple second gripper action candidates $\{g_2\}$. We then employ $\mathcal{C}_2$ to determine how well these second gripper candidates $\{g_2\}$ collaborate with the proposed g_1. The average critic score reflects how easily the second gripper can cooperate with the proposed first action g_1. Consequently, this average score serves as the ground truth for the first Critic Network $\mathcal{C}_1$, and we apply L1 loss for supervision:

$$\hat{c}_{g_1} = \frac{1}{nm}\sum_{j=1}^{n}\sum_{k=1}^{m} \mathcal{C}_2\big(f_{in}, f_{g_1}, f_{p_j}, \mathcal{U}_2(f_{in}, f_{g_1}, f_{p_j}, z_{jk})\big); \qquad \mathcal{L}_{\mathcal{C}_1} = \left|\mathcal{C}_1(f_{in}, f_{g_1}) - \hat{c}_{g_1}\right|. \tag{8}$$

To train the Affordance Network $\mathcal{A}_1$ and Actor Network $\mathcal{U}_1$ in $\mathcal{M}_1$, the loss functions are similar to those used for $\mathcal{A}_2$ and $\mathcal{U}_2$. Therefore, with the trained Critic Network $\mathcal{C}_1$, the Affordance Network $\mathcal{A}_1$ assigns high scores to points that can be easily manipulated collaboratively by the subsequent gripper action.

In this training pipeline, the two gripper modules can generate collaborative affordance maps and manipulation actions for bimanual tasks. Note that during inference, the use of the Critic Networks is optional.

D Details about Training and Computational Costs

During training, there are two main components: (1) the Disassembly Predictor and the Transformation Predictor are trained together in an end-to-end manner, and (2) all modules within the BiAssembly Predictor are also trained collectively in an end-to-end manner. These two training components can be conducted simultaneously on a single GPU. Using a single NVIDIA V100 GPU, the total training time for our model is approximately 48 hours: the combination of the Disassembly Predictor and Transformation Predictor converges in about 20 hours, while the BiAffordance Predictor converges in about 48 hours.

During inference, our method utilizes only 1,600 MB of GPU memory and processes each data point in an average of 0.1 seconds.

E Failure Cases

We provide a detailed analysis of failure cases and illustrate the inherent difficulty of the task with scenarios that are particularly challenging for robots to figure out. Additionally, we provide insights into potential future improvements to address these complexities more effectively.

E.1 Hard to Grasp

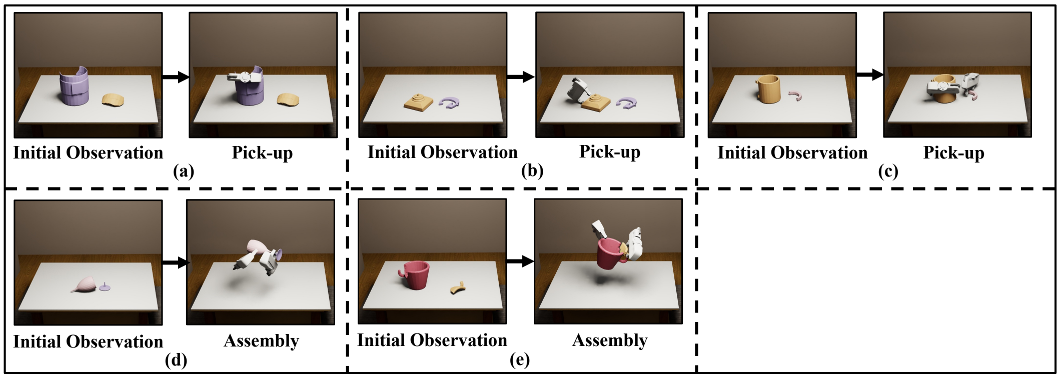

Figure 7: We visualize some failure cases, which demonstrate the challenges of the tasks and some cases that are difficult for robots to to determine appropriate actions. The first row presents three cases where the fractured parts are either too large or too flat to grasp. The second row includes two cases where the graspable region corresponds exactly to the seam areas; while the objects can be grasped, collisions may occur during the assembly of the parts.

Heavy or Smooth-Surfaced Parts. Fractured parts that are heavy or have smooth surfaces often result in grasping failures. For instance, as shown in Figure 7 (a), categories such as teapots and vases, which are relatively large and feature smooth curved surfaces, exhibit notably high failure rates during grasping.

Flat Parts. Flat fractured parts, particularly some shapes in categories like statues and mugs, are challenging to pick up due to the limited gripping area. For example, as shown in (b), the statue part on the left is too close to the desktop and has a very small thickness, which prevent the gripper from grasping it. Similarly, in (c), the handle fragment on the right is too flat, making it impossible fo rthe gripper to grasp it. A potential solution is incorporating pre-grasp operations, such as moving the fractured part to the table edge, allowing the shape to hang off slightly and thus become graspable.

E.2 Hard to Assemble

Graspable Regions Overlapping Seam Areas. When the graspable regions of a fractured part align with its seam areas, collisions during assembly become frequent. This issue is common in categories such as wineglasses, mugs, and bowls. For example, as shown in Figure 7 (d), the left gripper avoids collision-prone regions, but the right gripper must grasp the neck of the wine bottle. Similarly, in (e), while the left gripper avoids collisions, the right gripper ends up grasping the handle of a mug. A potential solution is to perform a series of pick-and-place operations to adjust the object's initial pose. This adjustment can reduce the overlap between the object's graspable regions and seam areas, thereby minimizing collisions during the assembly process.

Complex Object Shapes. Objects with intricate shapes, like those in the statues category, pose challenges due to irregular edges and complex curves. Such designs increase the difficulty of alignment and manipulation, leading to higher failure rates during assembly.

Relative Displacement During Operations. Relative displacement between the gripper and fractured parts often occurs due to small contact areas and insufficient support, which can cause sliding or tipping during manipulation. For example, wine bottles with narrow necks, which have unstable center of gravity, making the gripper prone to sliding during movement and leading to operational failures.

F Discussions and Future Works

F.1 Handling Multiple Broken Parts

Our method can be extended to handle multiple fragments. Below, we provide a detailed explanation of how our method can be adapted for multi-fragment assembly, followed by the experimental results.

The multi-fragment assembly task can be achieved by iteratively applying the two-fragment assembly process. First, at each iteration, we can identify which two fragments, $p_i$and p_j, should be assembled next. (If some parts have already been assembled in previous iterations, their combination is treated as a new fragment.) Specifically, based on the imaginary assembled shape S, we can calculate the minimum distance,$\min \|p_i - p_j\|$, between sampled points from every pair of fragments, and the pair $(p_i,\, p_j)$ with the minimum distance is chosen for assembly:$(p_i, p_j) = \underset{(p_i,p_j)\in\mathcal{S}_1\times\mathcal{S}_2}{\arg\min} \|p_i - p_j\|$.Once p_i and p_j are identified on S, we then map these fragments to their corresponding parts in the observed point cloud O. This mapping is formulated as a classification task, where the similarity between parts in S and O is estimated.

Finally, using the imaginary assembled shape of the selected fragments, $S_{p_i} \cup S_{p_j}$, and the corresponding observed point cloud $O_{p_i} \cup O_{p_j}$, our method predicts the actions to pick up and assemble the fragments. This process mirrors the steps of the standard two-fragment assembly method. By iteratively applying this two-fragment assembly process, the complete assembly of all fragments can be achieved. To validate the feasibility of this multi-fragment assembly process, we evaluated our pretrained BiAssembly model on broken beerbottles with three pieces without any finetune process. We provide the visualization of the predicted affordance maps and actions in Figure 8, we can see that for multi-fragment assembly task, our method can still predict reasonable results in each iteration.

While the above proposed method is a practical approach for assembling multi-part fractures, another potential strategy is training the Affordance Network to identify which two fragments are easiest to assemble in each iteration. In this new method, the Affordance Network would involve assigning high affordance scores to the reasonable regions of these fragments, while predicting low affordance scores for the fragments that are not being assembled in the current iteration. Implementing this strategy would require additional data collection for training and modifications to the framework. We leave this exploration for future work.

F.2 The Imaginary Assembled Shape

Predicting the imaginary assembled shape from multiple fractured parts is a relatively well-studied vision problem (Sellán et al., 2022; Wu et al., 2023c; Lu et al., 2024c; Tsesmelis et al., 2024; Scarpellini et al., 2024). Previous works have demonstrated the ability to predict precise fragment poses that allow for an imaginary assembled shape, making it reasonable to assume the existence of such shapes in our framework. Additionally, in traditional furniture assembly tasks, several studies (Wang et al., 2022a; Sera et al., 2021; Wan et al., 2024) also assume the existence of an imaginary assembled shape as part of their formulation. While this assumption aligns with advancements in prior works, we hope future research can achieve complex and challenging shape assembly tasks without depending on an imaginary assembled shape.

G More Experimental Results

In this section, we conduct three additional ablation studies, and provide the quantitative results in Table 4 and Table 5.

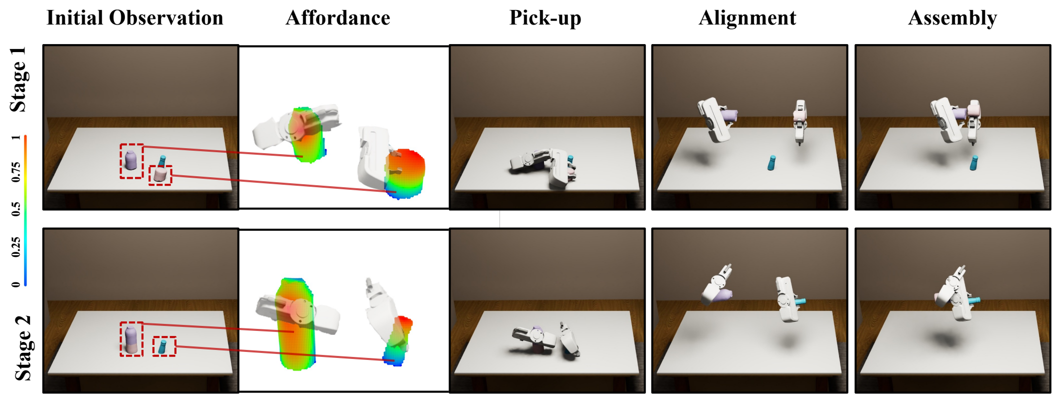

Figure 8: We provide the visualization of the predicted affordance maps and actions for multi-part assembly task.

w/o Affordance Network: During inference, we do not use the trained Affordance Network to highlight actionable regions. Instead, we randomly sample a contact point on the part. The results show a significant drop in the success rates, which decrease to 4.60% for training categories and 2.80% in unseen categories. This demonstrates that the Affordance Network plays a crucial role in filtering out non-graspable points and points that are unsuitable for the subsequent assembly process.

w/o Transformation Predictor: In this ablation, we remove the Transformation Predictor during inference. This results in success rates of 7.40% on training categories and 4.80% on unseen categories, both substantially lower than our original method. These results show that the Transformation Predictor plays an essential role in predicting alignment poses, enabling the robot to manipulate parts from their initial to alignment poses without collisions.

w/ heuristic disassembly direction v: In this case, we remove the Disassembly Predictor during inference. Instead, we compute the center of each part from the imaginary assembled shape S by averaging the part points, and then use the relative direction of the two parts' centers as the disassembly direction v. This ablation achieves success rates of 19.70% on training categories and 15.20% on unseen categories, both of which are lower than those achieved by our method. While this ablated version performs well on certain categories, suggesting that the calculated relative direction can approximate the relative positions of the two parts, it falls short in categories with complex geometries. In such cases, the heuristic method lacks the accuracy needed to replace the assembly direction required for our task. This highlights the critical role of the Disassembly Predictor in achieving better performance.

Table 4: More ablation studies: quantitative results in novel instances within training categories.

	Novel Instances in Training Categories										
Method											AVG
w/o Affordance	7%	11%	0%	0%	1%	8%	1%	4%	6%	8%	4.60%
w/o Transformation	29%	19%	0%	0%	0%	0%	8%	4%	5%	9%	7.40%
w/ heuristic v	54%	28%	0%	3%	10%	5%	**28%**	**23%**	21%	**25%**	19.70%
Ours	**60%**	**38%**	**13%**	**13%**	**12%**	**9%**	26%	18%	**27%**	**25%**	**24.10%**

Table 5: More Ablation studies: quantitative results in the novel unseen categories.

Method	Unseen Categories 					AVG
w/o Affordance	2%	6%	2%	0%	4%	2.8%
w/o Transformation	4%	10%	1%	0%	9%	4.8%
w/ heuristic v	**18%**	22%	**15%**	**9%**	12%	15.20%
Ours	14%	**31%**	10%	7%	**25%**	**17.4%**

H More Visualizations

In Figure 9, we present additional qualitative results from our method.

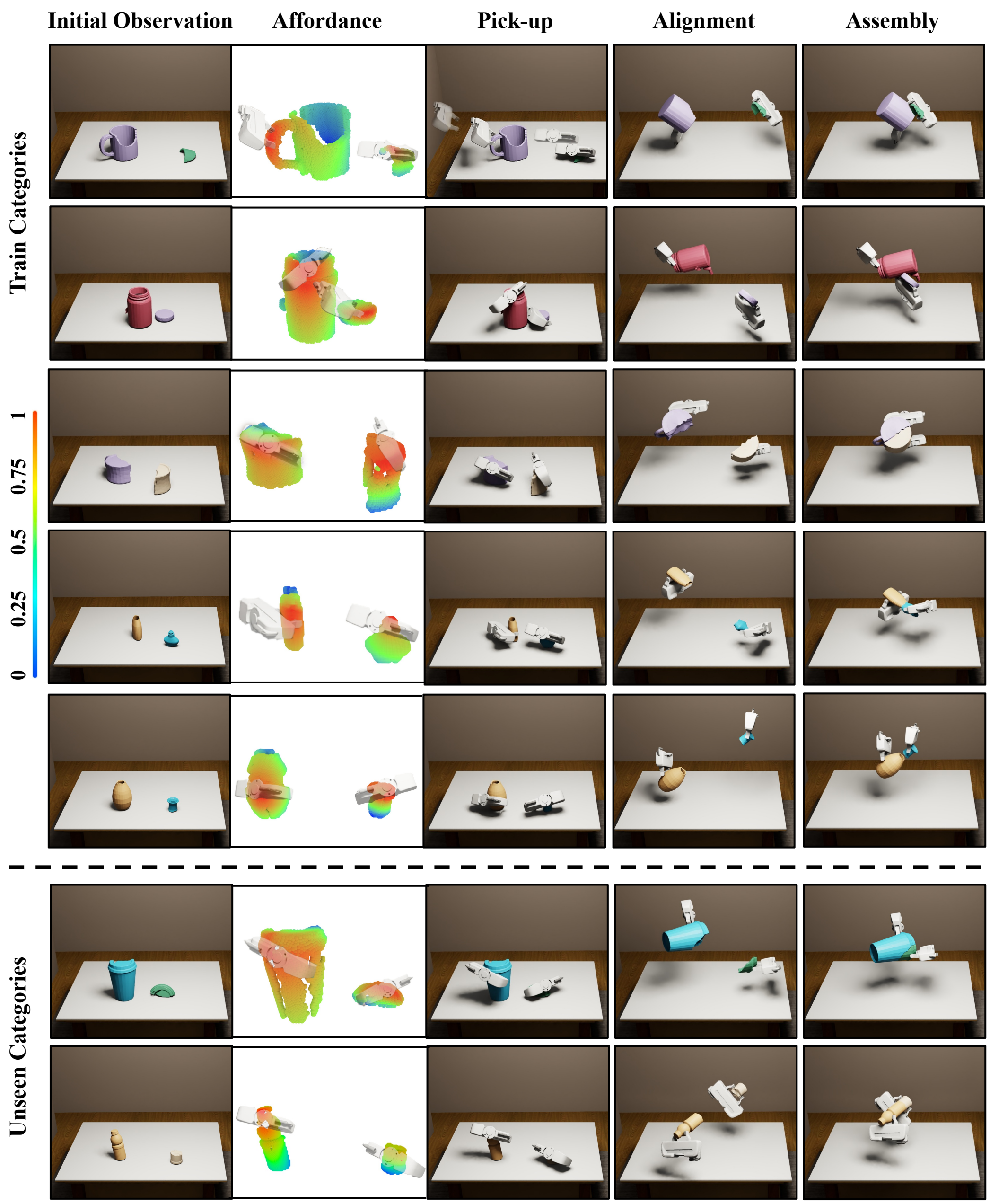

Figure 9: We visualize additional qualitative results that augment Figure 4 in the main paper. In each row, from left to right, we respectively present the input observation, the predicted affordance maps for the two fractured parts, and the bimanual actions for the pick-up, alignment, and assembly steps. In the top part are novel shapes from the training categories, while in the bottom part are shapes from unseen categories.