# SE(3)-Equivariant Diffusion Models for 3D Object Analysis

**Xie Min**, **Zhao Jieyu***, **Shen Kedi** and **Chen Kangxin**
Ningbo University
cqxiemin@gmail.com, zhao_jieyu@nbu.edu.cn, skuld_1456@163.com, chen_kangxin@outlook.com

## Abstract

SE(3)-equivariance is a critical property for capturing pose information in 3D vision tasks, enabling models to handle transformations such as rotations and translations effectively. While equivariant diffusion models have recently demonstrated promise in 3D object reassembly due to their generative and denoising capabilities, they face key challenges when applied to this task. Specifically, traditional diffusion models rely on fixed input sizes, which limits their adaptability to varying part quantities, and their linear noise addition and removal processes struggle to address the inherently nonlinear transformations of 3D parts. To overcome these limitations, this paper proposes an SE(3)-equivariant diffusion model for pose denoising and 3D object reassembly from fragmented parts. The model incorporates an equivariant encoder to extract SE(3)-equivariant features, a Lie algebra mapping to linearize noise addition and removal, and an elastic diffusion framework capable of adapting to varying part quantities and nonlinear transformations. By leveraging these components, the method achieves accurate and robust pose predictions across diverse input configurations. Experiments conducted on the Breaking Bad dataset, a real-world RePAIR and a self-constructed 3D mannequin dataset demonstrate the effectiveness of the proposed model, outperforming state-of-the-art methods across metrics such as root mean square error and part accuracy. Ablation studies further validate the critical contributions of key modules, emphasizing their roles in improving accuracy and robustness for 3D part reassembly tasks.

## 1 Introduction

3D vision is a crucial branch of computer vision that focuses on the comprehension and analysis of 3D objects, scenes, and motion. Humans instinctively solve complex spatial problems by leveraging spatial relationships and transformation equivariance to understand how various parts fit together. This capability is particularly valuable in fields such as archaeology, medicine, and biomedicine, where reconstructing fragmented 3D objects requires both accurate identity recognition and precise pose estimation. However, predicting accurately pose is a challenge for conventional invariant networks [Xie *et al.*, 2024], which produce constant outputs regardless of input transformations, lack the flexibility to capture spatial variability. This limitation has driven the development of advanced networks, such as equivariant networks, which can adapt to diverse transformations.

Equivariant networks ensure that a transformation applied to the input results in a correspondingly transformed output. This property facilitates the learning of part transformation and improves generalization to unseen data with various spatial configurations. Existing methods for achieving equivariance in 3D vision can be broadly categorized into equivariant kernel networks and equivariant tensor networks. Equivariant kernel networks [Hoogeboom *et al.*, 2022], [Hutchinson *et al.*, 2021] extend convolution kernels using higher-order representations, such as spherical harmonics or Lie group representations [Yu and Sun, 2024]. While equivariant tensor networks [Son *et al.*, 2024], [Assaad *et al.*, 2023], [Lei *et al.*, 2023] enrich scalar features by evolving them into tensor-valued representations, enhancing the understanding of input.

Despite these advances, predicting part poses of fragmented 3D objects remains challenging. Single-layer equivariant networks struggle to simultaneously capture fine-grained part-level pose information and assemble these parts into complete objects. To address this limitation, hierarchical structures, such as equivariant graph networks [Scarpellini *et al.*, 2024], have been proposed. These networks [Satorras *et al.*, 2021], [Le *et al.*, 2022], [Kofinas *et al.*, 2024], [Du *et al.*, 2022], [Meng *et al.*, 2024] use message-passing mechanisms to maintain spatial relationships and capture geometric relationships between parts. However, they often fall sort when parts undergo different transformations, such as rotations.

To overcome this challenge, we draw inspiration from diffusion models, which excel in handling uncertainty and iterative refinement. Diffusion models, successful in tasks such as molecule generation [Guan *et al.*, 2023], robotic manipulation [Ryu *et al.*, 2024] and trajectory prediction [Chen *et al.*, 2023], [Liu *et al.*, 2024], perform iterative denoising to recover information from noisy data. Unlike NeRF [Zhou *et al.*, 2025] or Gaussian splatting [Ma *et al.*, 2024], which fo-

*Corresponding author

cus on fixed viewpoints or scenes, diffusion models are well-suited to tasks requiring iterative refinement, such as 3D part pose learning. By treating random transformations as noise, we conceptualize the learning of transformed 3D part poses as a denoising process [Bansal *et al.*, 2024], where the goal is to recover original poses and assemble parts into objects.

Building on this intuition, we propose a novel framework that integrates diffusion models with SE(3)-equivariance, SE(3)-equivariance maintains consistency under rotations and translations in 3D Euclidean space. Equivariance ensures consistent transformations across parts, while diffusion models provide the flexibility to generate and refine poses in diverse spatial configurations. However, a key challenge is the fixed input size of conventional diffusion models, which limits their adaptability to varying part counts. Inspired by advancements in elastic diffusion for 2D images [Zheng *et al.*, 2024], [Haji-Ali *et al.*, 2024], we extend these principles to the 3D domain, enabling our model to process inputs with varying part quantities while maintaining equivariance.

We evaluate our approach on the public 3D part dataset Breaking Bad [Sellán *et al.*, 2022], a real-world archaeological dataset RePAIR [Tsesmelis *et al.*, 2024] and a self-built dataset of fragmented 3D mannequin. These datasets allow us to train geometry-based networks without relying on semantic labels, providing the robustness of our framework. In summary, we make the following contributions:

- We propose a novel Lie algebra mapping to linearize the processes of noise addition and removal, enabling the transformation of 3D input data into a representation compatible with diffusion models.
- Building on this, we design an elastic diffusion model capable of handling arbitrary numbers of 3D parts, effectively adapting to varying input sizes while accurately denoising and predicting the pose states.
- Comprehensive experiments conducted on the Breaking Bad dataset, real-world RePAIR dataset and a self-constructed 3D mannequin dataset demonstrate that our method significantly outperforms state-of-the-art approaches in terms of root mean square error, mean absolute error, and part accuracy for 3D part assembly tasks.

## 2 Related Works

### 2.1 Equivariant Networks

Our work focuses on SE(3)-equivariant networks, formally, an SE(3)-equivariant function $f$ satisfies $g(f(\mathcal{X})) = f(g(\mathcal{X}))$, for all $g \in \text{SE(3)}$ applied to an input $\mathcal{X}$. Group equivariant convolution extends traditional convolution to group convolution under discrete symmetric group actions. For instance, Thomas et al. [Thomas, 2019] and Fuchs et al. [Fuchs *et al.*, 2020] use spherical harmonics as high-order filters to design equivariant networks. However, these methods often lack interpretability and flexibility due to architectural constraints [Son *et al.*, 2024].

To overcome this limitation, recent works [Deng *et al.*, 2021], [Assaad *et al.*, 2023],[Lei *et al.*, 2023] extend scalar features to tensor-valued representations, enabling network to model complex transformations. Son et al.[Son *et al.*, 2024] further map 3D point coordinates into a high-dimensional sinusoidal feature spaces for enhanced shape compression.

However, these single-layer equivariant networks struggle with tasks requiring the integration of individual part poses into a cohesive structure. To address this, researchers are exploring hierarchical structures for equivariant networks that can better model these complex relationships.

### 2.2 Hierarchical Networks and Diffusion Models

Hierarchical graph structures have been proposed to model spatial relationships among parts. Graph neural networks, which are typically permutation-equivariant [Huang *et al.*, 2024], maintain topological relationships through message-passing mechanisms. For example, Kofinas et al. [Kofinas *et al.*, 2024] represent neural networks as hierarchical computational graphs, preserving equivariance and spatial structure.

Diffusion models have also been explored for hierarchical tasks. Gianluca et al. [Scarpellini *et al.*, 2024] propose a graph-based diffusion model for refining 3D part poses, treating fragments as nodes in a spatial graph. Similarly, Wang et al. [Wang *et al.*, 2024] extend the PuzzleFusion [Hossieni *et al.*, 2024] concept to 3D, iteratively refining 6-DoF alignment parameters. These models demonstrate the potential of diffusion-based approaches for iterative pose refinement.

However, conventional diffusion models are typically trained on fixed-size inputs, they struggle with varying aspect ratios during inference. Therefore, designing a flexible diffusion model to handle diverse input size is a new challenge.

### 2.3 Elastic Feature Selection

Recent advancements in 2D diffusion models address the challenge of variable input sizes [Zheng *et al.*, 2024], [Podell *et al.*, 2023]. For instance, ElasticDiffusion [Haji-Ali *et al.*, 2024] decouples global and local content generation, enabling robust synthesis across resolutions. Inspired by these methods, we design elastic diffusion models for 3D object analysis, allowing the network to adapt to varying part counts while preserving equivariance. This flexibility is crucial for assembly tasks, where input configurations can vary widely.

## 3 Equivariant Diffusion Models

This section introduces the proposed equivariant diffusion framework, which aims to denoise pose transformations and reassemble fragmented 3D parts into a coherent object. The framework consists of three main components: (1) an equivariant feature representation module to extract translational and rotational equivariant features, (2) a Lie algebra mapping to linearize the addition and removal of noise in the transformation matrices, (3) an elastic diffusion model designed to handle diverse input sizes while ensuring accurate denoising.

### 3.1 Equivariant Feature Representation

To obtain equivariant features for fragmented parts, we first translate the parts $\{P^k\}_{k=1}^K$ such that their centers lie at the origin, ensuring that the input parts $\{\bar{P}^k\}_{k=1}^K$ are translationally equivariant. Each part $P^k$ is a sampled point cloud containing 1000 points.

Then a rotation-equivariant encoder is employed to extract features and predict pose matrices $\{M_0^k\}_{k=1}^K$. In this paper, we utilize the encoder proposed in [Xie *et al.*, 2024], which leverages an equivariant masked autoencoder to learn robust features for 3D objects. This encoder's feature learning capability has been validated through its reconstruction performance on various datasets.

The predicted pose matrices are then input into the diffusion model. To disrupt the input, we replace Gaussian noise with transformed matrices during the forward process. During the reverse process, the diffusion model denoises the matrices to estimate the original transformations, enabling the reassembly of fragmented parts into a complete 3D object.

### 3.2 Lie Algebra Mapping

Directly replacing Gaussian noise with transformation matrices $\Phi_{noise}^k \in \mathbb{R}^{3\times3}$ often leads to degeneration in the object's dimensions, as shown in the first row (orange object) of Figure 1. This degeneration occurs because traditional diffusion models rely on linear noise addition/removal, which is incompatible with the nonlinear nature of transformation matrices.


Figure 1: Comparison of two different ways of adding noise. The first row is replacing the Gaussian noise directly with transformation matrix. The second row is mapping the process of adding noise into the Lie algebra space.

To address this issue, we map the transformation matrices to the Lie algebra space $\mathfrak{so}(3)$ [Batatia *et al.*, 2023], where nonlinear operations can be transformed into linear ones (as shown in the sencond row, the blue object, of Figure 1). Specifically, we use the logarithmic map $\log : \mathrm{SO}(3) \rightarrow \mathfrak{so}(3)$ to project rotation matrices $\Phi_{noise}^k$ into vectors $\phi_{noise}^k$:

$$\phi_{noise}^k = \log(\Phi_{noise}^k) = [\mathbf{u}\theta]_\times \tag{1}$$

$$\theta = \arccos \frac{tr(\Phi_{noise}^k) - 1}{2} \tag{2}$$

$$\mathbf{u} = \frac{(\Phi_{noise}^k - (\Phi_{noise}^k)^\top)^\vee}{2\sin\theta} \tag{3}$$

where $\mathbf{u} = [u_x, u_y, u_z]^\top$ is a unit vector representing the rotation axis, $\theta \in [0, \pi]$ is the rotation angle. $\phi_{noise}^k$ is a $3 \times 3$ skew-symmetric matrix, $(\cdot)^\vee$ is the mapping from a skew-symmetric matrix in $\mathfrak{so}(3)$ to a vector. The inverse mapping is performed via the exponential map $\exp : \mathfrak{so}(3) \rightarrow \mathrm{SO}(3)$:

$$\Phi_{noise}^k = \exp(\phi_{noise}^k) \tag{4}$$

By leveraging the Lie algebra mapping to linearize the noise addition and removal processes, the part pose features are transformed into a representation compatible with diffusion models. Building on this foundation, we propose an elastic diffusion model designed to handle diverse input sizes and perform effective denoising and pose prediction.

### 3.3 Elastic Diffusion Models

Before feeding these mapped vectors into diffusion model, we ensure compatibility with varying input sizes by applying padding and cropping operations. The padded vectors are:

$$\hat{\phi} \leftarrow f_P(\phi, \mathcal{Z}) \tag{5}$$

$$\dot{\phi} = f_C\left(\hat{\phi}, K\right) \tag{6}$$

where $\phi = \{\phi_{GT}, \phi_{noise}\}$, $\mathcal{Z}$ is a zero matrix for padding. Padding function $f_P$ ensures that the number of $\hat{\phi}$ is always $N$, cropping function $f_C$ ensures that the output size matches the original part count $K$. This padding-cropping strategy guarantees consistent input sizes $N$ while retaining the focus on the actual parts $K$.

The process of the diffusion model for 3D part pose learning is illustrated in Figure 2. Each row corresponds to the transformation and recovery of a single fragmented part. The framework consists of a forward diffusion process and a reverse diffusion process, which ensure accurate reconstruction of the original pose for each part, enabling effective 3D part reassembly.

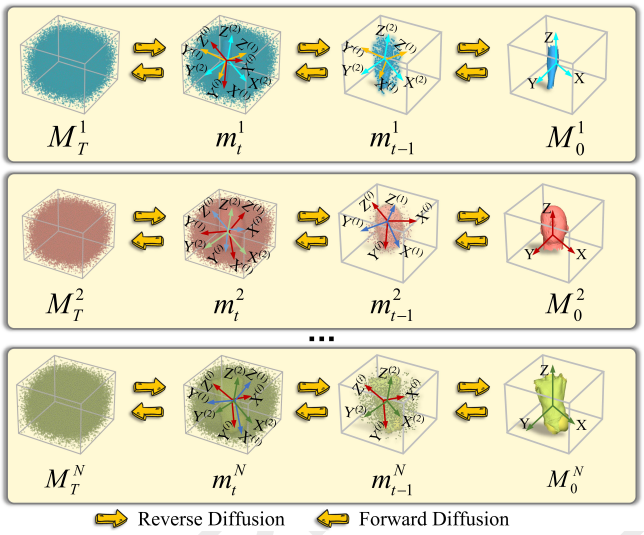


Figure 2: Illustration of the diffusion process for 3D part pose learning. Each row represents one part's pose transformation. In the forward diffusion process, transformation matrices progressively perturb the part pose, resulting in a noisy transformed state $M_T^k$. The reverse diffusion process iteratively removes the perturbations, denoising the pose to recover the original state $M_0^k$.

**Forward process.** In Lie algebra space, noise is added linearly to the transformed pose vectors:

$$\hat{m}_t^k = \hat{\phi}_{GT}^{k,t} + \hat{\phi}_{noise}^{k,t} \tag{7}$$

$$\hat{\phi}_{GT}^k = f_P(\log(M_{GT}^k), \mathcal{Z}) \tag{8}$$

$$\hat{\phi}_{noise}^k = f_P(\log(\Phi_{noise}^k), \mathcal{Z}) \tag{9}$$

where $\hat{m}_t^k$ represents the noisy pose at step $t$, $M_{GT}^k = M_0^k$ is the original pose state.

**Reverse process.** The model estimates and removes noise to recover the original pose:

$$\hat{m}_{t-1}^k = \hat{m}_t^k - \hat{\varphi}_{noise}^{k,t} \tag{10}$$

where $\hat{\varphi}_{noise}^{k,t}$ is the estimated noise output by our models that has to be removed from $\hat{m}_t^k$ at timestep $t$ to recover $\hat{m}_{t-1}^k$. The denoised Lie algebra element is then mapped back to SO(3) space:

$$\dot{M}_0^k = f_C(\exp(\hat{M}_0^k), K) \tag{11}$$

Finally, the denoised pose matrix $\dot{M}_0^k = \{\dot{R}_0^k, \dot{T}_0^k\}$ should correspond to the original pose matrix, ensuring accurate reassembly of the fragmented parts.

**Loss function.** To optimize the diffusion models, we introduce three loss functions: translation loss $\mathcal{L}_{trans}$, rotation loss $\mathcal{L}_{rot}$, and point loss $\mathcal{L}_{point}$. The total loss is:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{trans} + \lambda_2 \mathcal{L}_{rot} + \lambda_3 \mathcal{L}_{point} \tag{12}$$

where $\{\lambda_i\}_{i=1}^3$ are weights. Each term ensures precise reassemble of translation, rotation.

**Translation loss.** The translation loss computes the distance between the ground truth $\mathbf{T}_0$ and the predicted ones $\dot{\mathbf{T}}_0$:

$$\mathcal{L}_{trans} = \|\dot{\mathbf{T}}_0 - \mathbf{T}_0\|_2^2 \tag{13}$$

where $\mathbf{T}_0 = \{T_0^k\}_{k=1}^K$ is the ground truth of translation, $\|\cdot\|_2^2$ is the $\mathcal{L}2$ loss.

**Rotation loss.** The rotation loss measures the geodesic distance between $\dot{\mathbf{R}}_0$ and $\mathbf{R}_0$:

$$\mathcal{L}_{rot} = \arccos \frac{tr(\dot{\mathbf{R}}_0 \mathbf{R}_0^\top) - 1}{2} \tag{14}$$

**Point loss.** We further use Chamfer Distance $f_{cd}(\cdot)$ to jointly measure the difference by supervising the reassembled pose of point cloud:

$$\mathcal{L}_{point} = f_{cd}(\mathbf{P}\dot{\mathbf{R}}_0 + \dot{\mathbf{T}}_0, \mathbf{P}\mathbf{R}_0 + \mathbf{T}_0) \tag{15}$$

where $\mathbf{P} = \{\bar{P}^k\}_{k=1}^K$ is the input parts.

# 4 Experiments

## 4.1 Dataset and Implementation Details

**Datasets.** This study focuses on denoising poses to recover the original part states and reassembling transformed parts into a coherent object. To evaluate the proposed approach, we conduct experiments on the publicly available Breaking Bad dataset [Sellán *et al.*, 2022], a real-world RePAIR dataset [Tsesmelis *et al.*, 2024] and construct a self-built 3D mannequin dataset to assess the robustness of our method.

The Breaking Bad dataset [Sellán *et al.*, 2022] consists of approximately 10,000 meshes derived from PartNet [Mo *et al.*, 2019] and Thingi10k [Zhou *et al.*, 2016]. Specifically, we select samples from the *everyday* subset and 6 *artifact* categories (sculpture, spiral bulbs, rabbit, frog, sofa and boy) for experiments. The *everyday* subset contains 20 object categories, each containing fragments with varying quantities. The RePAIR dataset [Tsesmelis *et al.*, 2024] consists of over 1,000 verified fragmented parts, it contains detailed 3D scans of the fragments, the fragments and fractures are realistic, caused by a collapse of a fresco during a World War II bombing at the Pompeii archaeological park. While the Breaking Bad and RePAIR datasets provides diverse object categories, it lacks sufficient variation in human-like object structures, which are essential for evaluating real-world applications such as pose estimation and archaeological human fragment restoration.

To address this limitation, we construct the 3D mannequin dataset, focusing on human-like structures with predefined semantic parts. This dataset consists of 133 simulated mannequin samples collected from ShapeNet [Yi *et al.*, 2016]. These samples are preprocessed by SubdivNet [Hu *et al.*, 2022] to ensure smoothness and uniformity, resulting in mesh with 16,384 faces each. The mannequins are segmented into distinct fragments based on semantic labels, with each fragment stored in OBJ format. We also create five subsets with varying fragment proportions. For example, a sample in the Mannequin_4Parts subset includes a head, torso, arms, and legs, as illustrated in Figure 3.

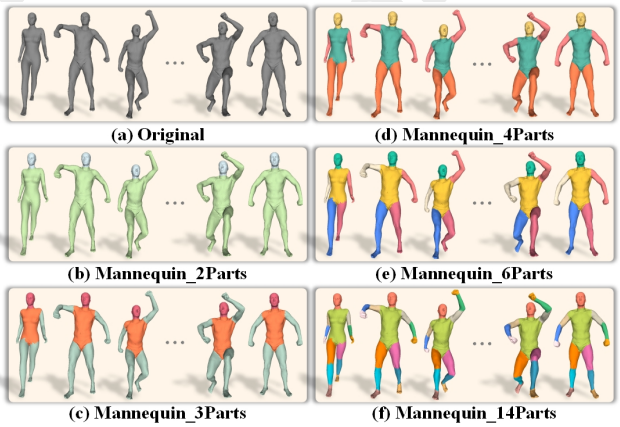


Figure 3: 3D Mannequin_$n$Parts dataets, $n = 2, 3, 4, 6, 14$. The segment parts of them are derived from semantic labels.

**Implementation details.** All experiments are performed on a Linux workstation equipped with an NVIDIA RTX 4090 GPU. We use Noise Conditional Score Networks (NCSNs) [Song and Ermon, 2019], [Croitoru *et al.*, 2023] as the backbone of the diffusion models. For the 3D object reassembly task, we adopt the encoder proposed in [Xie *et al.*, 2024] to extract features from each part and predict the corresponding poses. These predicted poses are then fed into the diffusion models for denoising, allowing the refinement of poses toward their original states.

Optimization is performed using the Adam optimizer with an initial learning rate of 1e-3 and a cosine learning schedule. The batch size is set to 16, and the total number of iterations is 2000, with a consistent input size $N = 30$. To further improve performance, we incorporate a Chamfer Distance loss term, as proposed in [Sellán *et al.*, 2022], which enhances the model's capability to minimize geometric discrepancies.

**Evaluation metrics**. We evaluate our method using multi-

ple metrics, including root mean square error (RMSE) metrics [Sellán *et al.*, 2022], RMSE(R) for rotation and RMSE(T) for translation, and the mean absolute error (MAE) for both rotation and translation. Additionally, part accuracy (PA) [Sellán *et al.*, 2022] employed to assess the precision of parts.

### 4.2 3D Object Reassembly Task

The quantitative evaluation of our method, compared to state-of-the-art methods, on the 3D Mannequin_$n$Parts datasets is presented in Table 1. The results indicate that while the performance of our method decreases with an increasing number of parts, it consistently ourperforms SE(3)-Equiv [Wu *et al.*, 2023], Jigsaw [Lu *et al.*, 2024] and DiffAssemble [Scarpellini *et al.*, 2024]. Visualization results are shown in Figure 4, and the iterative refinement process is illustrated in Figure 6.

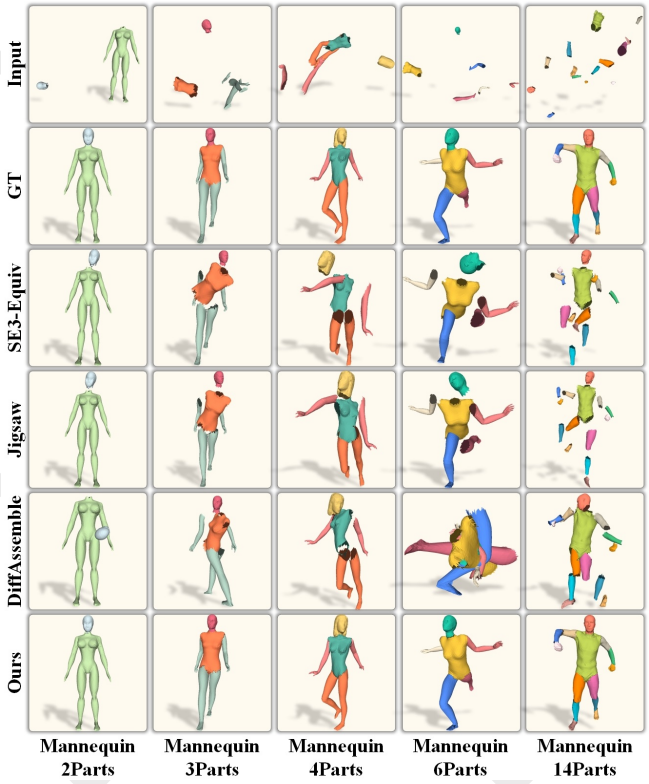


Figure 4: Qualitative results on 3D Mannequin_$n$Parts datasets for various multi-geometric assembly. The input is transformed mannequin parts.

Qualitative results demonstrate that our method effectively denoises the pose noise and iteratively predicts the original poses. Notably, we observe that SE(3)-Equiv and Jigsaw struggle with larger parts, while DiffAssemble tends to shift all parts toward the center, as shown in Figure 4. It is worth noting that the runtime of the method proposed in this article is measured from the moment the obtained poses are fed into the diffusion model. This is primarily done to evaluate the diffusion model's ability to learn poses.

We also evaluate our method on the Breaking Bad and RePAIR datasets, we select all categories from the *everyday* subset and 6 categories from *artifact* subset of Breaking Bad dataset as experimental subjects. Quantitative results are summarized in Table 2 and 3.

These results highlight superior performance of our method across various metrics, emphasizing its accuracy and robustness in 3D object reassembly task. Although our method requires more time compared to SE(3)-Equiv [Wu *et al.*, 2023], CCS [Zhang *et al.*, 2024] and Jigsaw [Lu *et al.*, 2024], its superior performance justifies the additional computational cost. Visualization results are shown in Figure 6, and the iterative process is illustrated in Figure 7.

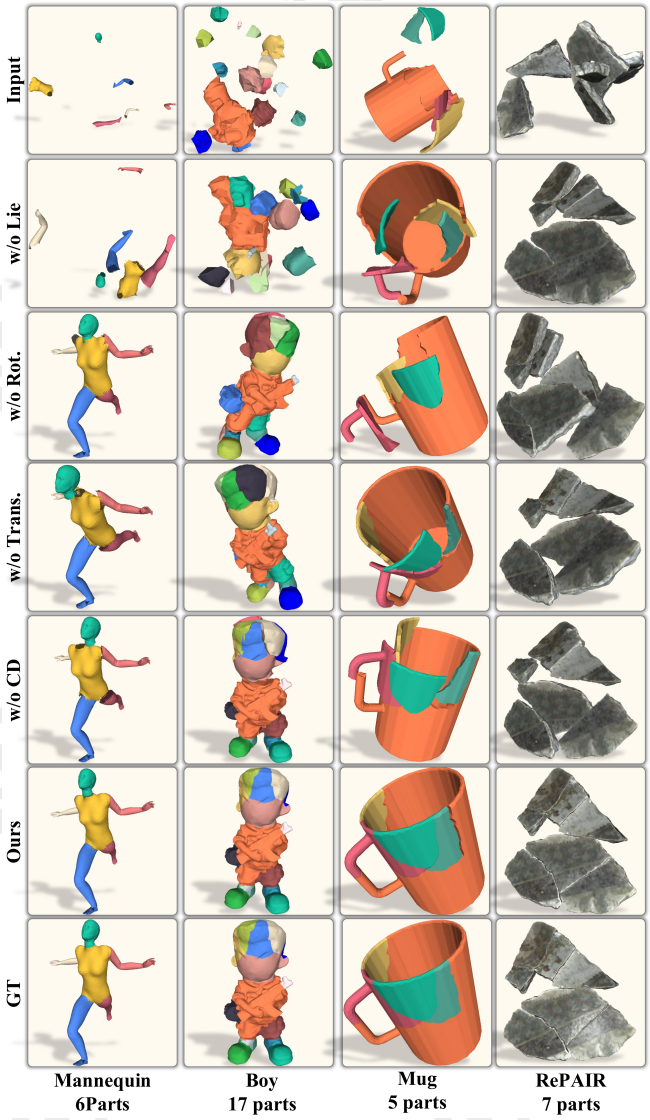


Figure 5: Ablation visual results on Mannequin_6Parts, Breaking Bad and RePAIR datasets.

### 4.3 Ablation Studies

To evaluate the contribution of different modules, we conduct ablation studies on Mannequin_6Parts, Breaking Bad and RePAIR datasets. **w/o Lie:** without mapping input and noise into the Lie algebra space, **w/o Rot.:** without mapping ro-

| Methods | Dataset | RMSE(R)↓ (degree) | MAE(R)↓ (degree) | RMSE(T)↓ $\times 10^{-2}$ | MAE(T)↓ $\times 10^{-2}$ | PA↑ (%) | Times (s/sample) |
|---|---|---|---|---|---|---|---|
| SE(3)-Equiv [Wu *et al.*, 2023] | Mannequin_2Parts | 79.8 | 69.3 | 21.5 | 19.0 | 16.7 | 0.72 |
| | Mannequin_3Parts | 89.7 | 76.5 | 12.0 | 10.5 | 24.8 | 0.78 |
| | Mannequin_4Parts | 88.8 | 77.2 | 12.7 | 10.9 | 15.4 | 0.77 |
| | Mannequin_6Parts | 88.9 | 77.5 | 20.9 | 17.9 | 18.4 | 0.70 |
| | Mannequin_14Parts | 86.4 | 74.4 | 13.6 | 11.6 | 13.9 | 1.59 |
| Jigsaw [Lu *et al.*, 2024] | Mannequin_2Parts | 57.6 | 49.8 | 9.5 | 6.9 | 33.3 | 0.71 |
| | Mannequin_3Parts | 57.6 | 50.0 | 9.8 | 7.2 | 33.3 | 0.45 |
| | Mannequin_4Parts | 62.7 | 54.3 | 20.1 | 15.9 | 25.6 | 0.55 |
| | Mannequin_6Parts | 66.4 | 57.8 | 13.9 | 11.9 | 23.1 | 0.33 |
| | Mannequin_14Parts | 80.0 | 69.8 | 15.0 | 12.7 | 12.8 | 1.81 |
| DiffAssemble [Scarpellini *et al.*, 2024] | Mannequin_2Parts | 81.3 | - | 27.1 | - | 5.1 | 0.85 |
| | Mannequin_3Parts | 67.2 | - | 12.7 | - | 26.7 | 0.86 |
| | Mannequin_4Parts | 72.5 | - | 17.5 | - | 11.1 | 1.54 |
| | Mannequin_6Parts | 82.1 | - | 25.5 | - | 13.2 | 1.21 |
| | Mannequin_14Parts | 75.1 | - | 12.4 | - | 32.2 | 1.07 |
| Ours | Mannequin_2Parts | 22.3 | 19.8 | 3.9 | 2.6 | 74.4 | 0.75 |
| | Mannequin_3Parts | 19.4 | 18.1 | 10.6 | 10.0 | 67.0 | 0.76 |
| | Mannequin_4Parts | 22.4 | 16.0 | 6.0 | 5.3 | 63.8 | 0.76 |
| | Mannequin_6Parts | 19.1 | 16.6 | 6.0 | 5.1 | 59.4 | 0.75 |
| | Mannequin_14Parts | 15.8 | 14.1 | 10.9 | 5.6 | 59.0 | 0.76 |

Table 1: Quantitative results on 3D Mannequin_$n$Parts datasets ($n = 2, 3, 4, 6, 14$) for various multi-geometric assembly.

| Methods | RMSE(R)↓ (degree) | MAE(R)↓ (degree) | RMSE(T)↓ $\times 10^{-2}$ | MAE(T)↓ $\times 10^{-2}$ | PA↑ (%) | Times (s/sample) |
|---|---|---|---|---|---|---|
| Global [Zhan *et al.*, 2020] | 79.2 | 66.3 | 14.7 | 11.7 | 23.1 | 0.45 |
| DGL [Zhan *et al.*, 2020] | 80.6 | 67.8 | 15.8 | 12.5 | 23.9 | 1.36 |
| RGL [Narayan *et al.*, 2022] | 83.2 | 70.8 | 14.9 | 11.8 | 25.4 | 0.60 |
| LSTM [Zhan *et al.*, 2020] | 83.0 | 71.0 | 15.3 | 12.1 | 21.7 | 1.28 |
| SE(3)-Equiv [Wu *et al.*, 2023] | 79.7 | 66.8 | 16.2 | 12.4 | 13.5 | 0.12 |
| CCS [Zhang *et al.*, 2024] | 85.0 | 74.4 | 13.4 | 8.9 | 13.4 | 0.41 |
| Jigsaw [Lu *et al.*, 2024] | 80.8 | 70.1 | 14.5 | 11.5 | 27.6 | 0.38 |
| DiffAssemble [Scarpellini *et al.*, 2024] | 73.3 | - | 14.8 | - | 27.5 | - |
| Ours | 18.3 | 15.7 | 4.1 | 3.1 | 59.7 | 0.73 |

Table 2: Quantitative results on Breaking Bad dataset (*everyday* subset) for multi-geometric assembly.

| Dataset | | Part num. | RMSE(R)↓ (degree) | MAE(R)↓ (degree) | RMSE(T)↓ $\times 10^{-2}$ | MAE(T)↓ $\times 10^{-2}$ | PA↑ (%) | Times (s/sample) |
|---|---|---|---|---|---|---|---|---|
| *artifact* subset | Sculpture | 3 | 11.5 | 9.9 | 2.8 | 2.7 | 53.9 | 0.75 |
| | Spiral bulbs | 7 | 19.6 | 18.9 | 2.8 | 2.4 | 53.7 | 0.55 |
| | Rabbit | 9 | 20.2 | 18.2 | 2.1 | 1.8 | 45.4 | 0.53 |
| | Frog | 15 | 29.0 | 24.1 | 3.2 | 3.1 | 65.7 | 0.54 |
| | Sofa | 15 | 20.2 | 14.3 | 4.0 | 3.2 | 46.7 | 0.76 |
| | Boy | 15 | 16.0 | 13.8 | 2.3 | 2.3 | 60.1 | 0.57 |
| RePAIR [Tsesmelis *et al.*, 2024] | | - | 17.7 | 12.9 | 5.4 | 4.8 | 59.9 | 0.77 |

Table 3: Quantitative results on Breaking Bad (*artifact* subset) and RePAIR datasets for multi-geometric assembly.

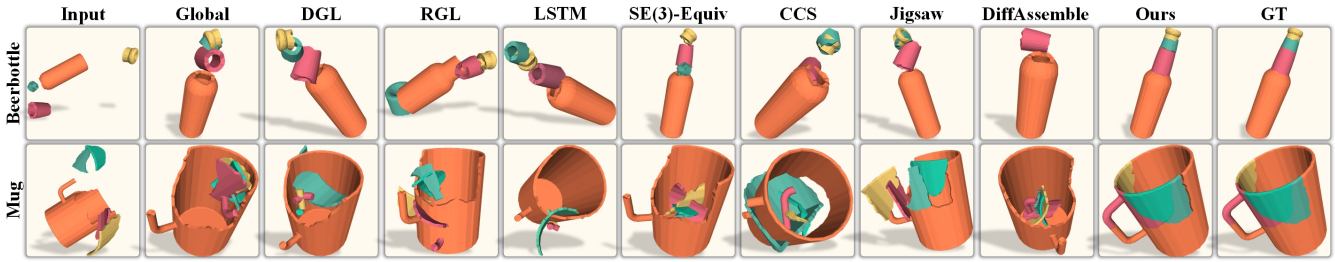


Figure 6: Qualitative results on Breaking Bad dataset (*everyday* subset).

tation noise into the Lie algebra space, **w/o Trans.:** without mapping translation noise into the Lie algebra space, **w/o CD:** without the Chamfer Distance as an additional test metric.

The results (Table 4 and Figure 5) show a significant performance decline when the Lie algebra mapping is removed. This decline underscores the critical role of Lie algebra mapping in linearizing the noise addition and denoising processes. Furthermore, Figure 5 (second column) reveals that the absence of Lie mapping leads to disorganized pose predictions, as the model fails to effectively disentangle pose noise.

The impact of removing rotation noise mapping is particularly pronounced. Without this step, the network struggles

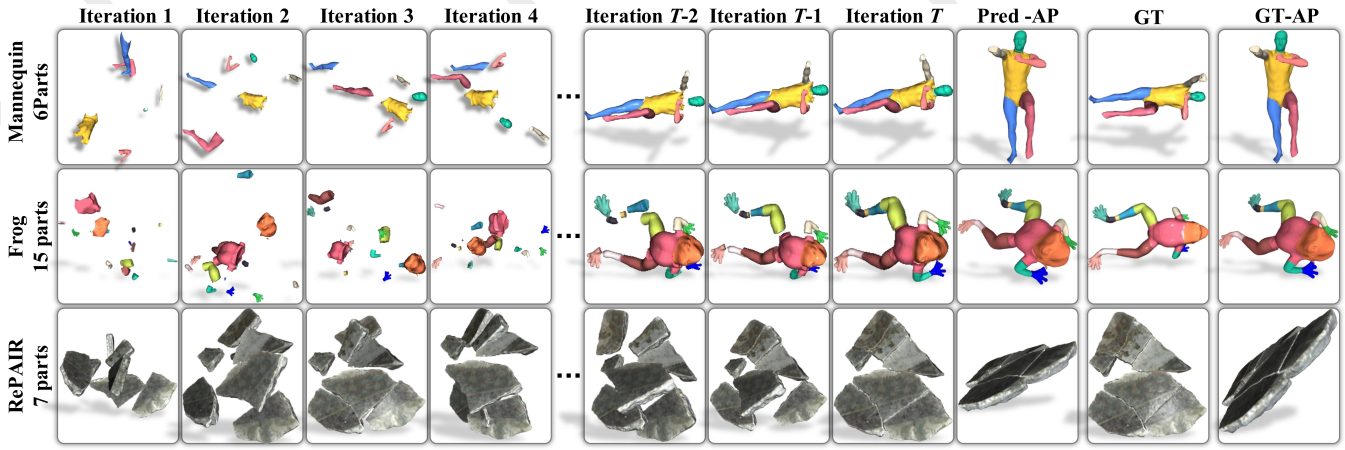


Figure 7: Visualization of the assembly process in the iterations on Mannequin_6Parts, Breaking Bad and RePAIR datasets. Pred-AP and GT-AP are the another perspective of the prediction and the ground truth, respectively.

| Dataset | Methods | RMSE(R)↓ (degree) | MAE(R)↓ (degree) | RMSE(T)↓ $\times 10^{-2}$ | MAE(T)↓ $\times 10^{-2}$ | PA↑ (%) | Times (s/sample) |
|---|---|---|---|---|---|---|---|
| Mannequin_6Parts | w/o Lie | 69.5 | 65.5 | 35.1 | 30.0 | 25.5 | 0.28 |
| | w/o Rot. | 29.5 | 40.1 | 7.2 | 7.4 | 49.7 | 0.66 |
| | w/o Trans. | 19.1 | 16.7 | 11.4 | 10.6 | 52.1 | 0.70 |
| | w/o CD | 19.5 | 40.1 | 6.3 | 5.2 | 51.4 | 0.61 |
| | Ours | 19.1 | 16.6 | 6.0 | 5.1 | 59.4 | 0.75 |
| *everyday* subset | w/o Lie | 44.2 | 54.3 | 24.5 | 21.7 | 16.5 | 0.27 |
| | w/o Rot. | 19.4 | 17.1 | 6.3 | 5.1 | 43.3 | 0.64 |
| | w/o Trans. | 18.7 | 16.1 | 5.9 | 5.2 | 46.0 | 0.68 |
| | w/o CD | 19.5 | 16.2 | 5.9 | 5.5 | 38.2 | 0.65 |
| | Ours | 18.3 | 15.7 | 4.1 | 3.1 | 59.7 | 0.73 |
| RePAIR | w/o Lie | 47.9 | 27.5 | 8.1 | 5.2 | 40.4 | 0.33 |
| | w/o Rot. | 31.3 | 34.1 | 5.9 | 5.0 | 51.6 | 0.50 |
| | w/o Trans. | 21.2 | 15.2 | 6.1 | 5.7 | 54.9 | 0.62 |
| | w/o CD | 16.6 | 13.3 | 6.1 | 4.1 | 55.8 | 0.68 |
| | Ours | 17.7 | 12.9 | 5.4 | 4.8 | 59.9 | 0.77 |

Table 4: Ablations on Mannequin_6Parts, Breaking Bad and RePAIR datasets.

to learn pose features due to the inherent nonlinearity and complexity of rotational transformations. The visualization results further confirm this observation, showing erratic and inconsistent pose predictions for parts subjected to rotations.

The removal of translation noise mapping has a relatively smaller impact on the overall performance, though it still introduces noticeable errors in translation estimation. This suggests that while both rotation and translation mappings are important, the rotational component plays a more critical role in ensuring robust pose predictions.

The inclusion of Chamfer Distance loss improves part accuracy by refining the geometric alignment between predicted and ground truth poses. It encourages the model to prioritize local geometric alignment over global pose accuracy.

## 5 Conclusion and Limitation

This paper presents an SE(3)-equivariant diffusion model for predicting transformed part poses and reassembling fragmented parts into a complete object. The framework consists of three main components: an equivariant feature encoder, a Lie algebra mapping module and an elastic diffusion model. The encoder extracts rotational and translational equivariant features, the Lie algebra mapping enables the model to handle transformations in a linearized manner, and the elastic diffusion model leverages NCSNs with an elastic property, allowing the framework to adapt to varying part quantities. These design elements collectively enhance the model's robustness and adaptability for complex part-based reassembly tasks.

**Limitations and future work.** The current method is limited to rigid transformations, such as rotation and translation, which restrict its application to scenarios requiring flexible transformations, such as sliding, bending, or stretching. Incorporating flexible transformations without disassembling parts for forward prediction remains a challenging yet essential direction for improvement. Future work will focus on extending the model to handle non-rigid deformations by integrating techniques that can capture local pose changes driven by global structural deformations. This extension would enable the model to predict transformations for objects with complex, flexible morphologies, broadening its applicability to domains such as biomechanical modeling, soft robotics, and human body pose estimation.

## Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant 62471266, as well as Ningbo Major Research and Development Plan under Grant 2023Z224.

## References

[Assaad *et al.*, 2023] Serge Assaad, Carlton Downey, Rami Al-Rfou, Nigamaa Nayakanti, and Benjamin Sapp. Vn-transformer: Rotation-equivariant attention for vector neurons. *Transactions on Machine Learning Research*, 2023.

[Bansal *et al.*, 2024] Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise. *Advances in Neural Information Processing Systems*, 36, 2024.

[Batatia *et al.*, 2023] Ilyes Batatia, Mario Geiger, Jose Munoz, Tess Smidt, Lior Silberman, and Christoph Ortner. A general framework for equivariant neural networks on reductive lie groups. *Advances in Neural Information Processing Systems*, 36:55260–55284, 2023.

[Chen *et al.*, 2023] Kehua Chen, Xianda Chen, Zihan Yu, Meixin Zhu, and Hai Yang. Equidiff: A conditional equivariant diffusion model for trajectory prediction. In *International Conference on Intelligent Transportation Systems*, pages 746–751, 2023.

[Croitoru *et al.*, 2023] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023.

[Deng *et al.*, 2021] 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 *IEEE International Conference on Computer Vision*, pages 12200–12209, 2021.

[Du *et al.*, 2022] Weitao Du, He Zhang, Yuanqi Du, Qi Meng, Wei Chen, Nanning Zheng, Bin Shao, and Tie-Yan Liu. Se (3) equivariant graph neural networks with complete local frames. In *International Conference on Machine Learning*, pages 5583–5608, 2022.

[Fuchs *et al.*, 2020] Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se (3)-transformers: 3d roto-translation equivariant attention networks. *Advances in Neural Information Processing Systems*, 33:1970–1981, 2020.

[Guan *et al.*, 2023] Jiaqi Guan, Wesley Wei Qian, Xingang Peng, Yufeng Su, Jian Peng, and Jianzhu Ma. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. In *International Conference on Learning Representations*, 2023.

[Haji-Ali *et al.*, 2024] Moayed Haji-Ali, Guha Balakrishnan, and Vicente Ordonez. Elasticdiffusion: Training-free arbitrary size image generation through global-local content separation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6603–6612, 2024.

[Hoogeboom *et al.*, 2022] Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International Conference on Machine Learning*, pages 8867–8887, 2022.

[Hossieni *et al.*, 2024] Sepidehsadat Sepid Hossieni, Mohammad Amin Shabani, Saghar Irandoust, and Yasutaka Furukawa. Puzzlefusion: unleashing the power of diffusion models for spatial puzzle solving. *Advances in Neural Information Processing Systems*, 36, 2024.

[Hu *et al.*, 2022] Shi-Min Hu, Zheng-Ning Liu, Meng-Hao Guo, Jun-Xiong Cai, Jiahui Huang, Tai-Jiang Mu, and Ralph R Martin. Subdivision-based mesh convolution networks. *ACM Transactions on Graphics*, 41(3):1–16, 2022.

[Huang *et al.*, 2024] Ningyuan Huang, Ron Levie, and Soledad Villar. Approximately equivariant graph networks. *Advances in Neural Information Processing Systems*, 36, 2024.

[Hutchinson *et al.*, 2021] Michael J Hutchinson, Charline Le Lan, Sheheryar Zaidi, Emilien Dupont, Yee Whye Teh, and Hyunjik Kim. Lietransformer: Equivariant self-attention for lie groups. In *International Conference on Machine Learning*, pages 4533–4543, 2021.

[Kofinas *et al.*, 2024] Miltiadis Kofinas, Boris Knyazev, Yan Zhang, Yunlu Chen, Gertjan J Burghouts, Efstratios Gavves, Cees GM Snoek, and David W Zhang. Graph neural networks for learning equivariant representations of neural networks. In *International Conference on Learning Representations*, 2024.

[Le *et al.*, 2022] Tuan Le, Frank Noé, and Djork-Arné Clevert. Equivariant graph attention networks for molecular property prediction. *arXiv preprint arXiv:2202.09891*, 2022.

[Lei *et al.*, 2023] Jiahui Lei, Congyue Deng, Karl Schmeckpeper, Leonidas Guibas, and Kostas Daniilidis. Efem: Equivariant neural field expectation maximization for 3d object segmentation without scene supervision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4902–4912, 2023.

[Liu *et al.*, 2024] Yanghong Liu, Xingping Dong, Yutian Lin, and Mang Ye. Diftraj: Diffusion inspired by intrinsic intention and extrinsic interaction for multi-modal trajectory prediction. In *International Joint Conference on Artificial Intelligence*, pages 1–9, 2024.

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

[Ma *et al.*, 2024] Yikun Ma, Dandan Zhan, and Zhi Jin. Fastscene: Text-driven fast 3d indoor scene generation via panoramic gaussian splatting. In *International Joint Conference on Artificial Intelligence*, pages 1–9, 2024.

[Meng *et al.*, 2024] Ziqiao Meng, Liang Zeng, Zixing Song, Tingyang Xu, Peilin Zhao, and Irwin King. Towards geometric normalization techniques in se (3) equivariant graph neural networks for physical dynamics simulations. In *International Joint Conference on Artificial Intelligence*, pages 1–9, 2024.

[Mo *et al.*, 2019] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su.

Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 909–918, 2019.

[Narayan *et al.*, 2022] Abhinav Narayan, Rajendra Nagar, and Shanmuganathan Raman. Rgl-net: A recurrent graph learning framework for progressive part assembly. In *IEEE Winter Conference on Applications of Computer Vision*, pages 78–87, 2022.

[Podell *et al.*, 2023] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[Ryu *et al.*, 2024] Hyunwoo Ryu, Jiwoo Kim, Hyunseok An, Junwoo Chang, Joohwan Seo, Taehan Kim, Yubin Kim, Chaewon Hwang, Jongeun Choi, and Roberto Horowitz. Diffusion-edfs: Bi-equivariant denoising generative modeling on se (3) for visual robotic manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 18007–18018, 2024.

[Satorras *et al.*, 2021] Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International Conference on Machine Learning*, pages 9323–9332, 2021.

[Scarpellini *et al.*, 2024] Gianluca Scarpellini, Stefano Fiorini, Francesco Giuliari, Pietro Moreiro, and Alessio Del Bue. Diffassemble: A unified graph-diffusion model for 2d and 3d reassembly. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 28098–28108, 2024.

[Sellán *et al.*, 2022] Silvia Sellán, Yun-Chun Chen, Ziyi Wu, Animesh Garg, and Alec Jacobson. Breaking bad: A dataset for geometric fracture and reassembly. *Advances in Neural Information Processing Systems*, 35:38885–38898, 2022.

[Son *et al.*, 2024] Dongwon Son, Jaehyung Kim, Sanghyeon Son, and Beomjoon Kim. An intuitive multi-frequency feature representation for so (3)-equivariant networks. In *International Conference on Learning Representations*, 2024.

[Song and Ermon, 2019] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32:1–13, 2019.

[Thomas, 2019] Nathaniel Cabot Thomas. *Euclidean-equivariant functions on three-dimensional point clouds*. Stanford University, 2019.

[Tsesmelis *et al.*, 2024] Theodore Tsesmelis, Luca Palmieri, Marina Khoroshiltseva, Adeela Islam, Gur Elkin, Ofir I 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. *Advances in Neural Information Processing Systems*, 37:30076–30105, 2024.

[Wang *et al.*, 2024] Zhengqing Wang, Jiacheng Chen, and Yasutaka Furukawa. Puzzlefusion++: Auto-agglomerative 3d fracture assembly by denoise and verify. *arXiv preprint arXiv:2406.00259*, 2024.

[Wu *et al.*, 2023] Ruihai Wu, Chenrui Tie, Yushi Du, Yan Zhao, and Hao Dong. Leveraging se (3) equivariance for learning 3d geometric shape assembly. In *IEEE International Conference on Computer Vision*, pages 14311–14320, 2023.

[Xie *et al.*, 2024] Min Xie, Jieyu Zhao, and Kedi Shen. A novel so (3) rotational equivariant masked autoencoder for 3d mesh object analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–14, 2024.

[Yi *et al.*, 2016] Li Yi, Vladimir G Kim, Duygu Ceylan, I-Chao Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, and Leonidas Guibas. A scalable active framework for region annotation in 3d shape collections. *ACM Transactions on Graphics*, 35(6):1–12, 2016.

[Yu and Sun, 2024] Ruixuan Yu and Jian Sun. Pose-transformed equivariant network for 3d point trajectory prediction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512, 2024.

[Zhan *et al.*, 2020] 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.

[Zhang *et al.*, 2024] Ruiyuan Zhang, Jiaxiang Liu, Zexi Li, Hao Dong, Jie Fu, and Chao Wu. Scalable geometric fracture assembly via co-creation space among assemblers. In *AAAI Conference on Artificial Intelligence*, volume 38, pages 7269–7277, 2024.

[Zheng *et al.*, 2024] Qingping Zheng, Yuanfan Guo, Jiankang Deng, Jianhua Han, Ying Li, Songcen Xu, and Hang Xu. Any-size-diffusion: Toward efficient text-driven synthesis for any-size hd images. In *AAAI Conference on Artificial Intelligence*, volume 38, pages 7571–7578, 2024.

[Zhou *et al.*, 2016] Qingnan Zhou, Eitan Grinspun, Denis Zorin, and Alec Jacobson. Mesh arrangements for solid geometry. *ACM Transactions on Graphics*, 35(4):1–15, 2016.

[Zhou *et al.*, 2025] Qunjie Zhou, Maxim Maximov, Or Litany, and Laura Leal-Taixé. The nerfect match: Exploring nerf features for visual localization. In *European Conference on Computer Vision*, pages 108–127, 2025.