GARF: Learning Generalizable 3D Reassembly for Real-World Fractures

Sihang Li[1,*], Zeyu Jiang[1,*], Grace Chen[1,†], Chenyang Xu[1,†], Siqi Tan[1], Xue Wang[1], Irving Fang[1], Kristof Zyskowski[2], Shannon P. McPherron[3], Radu Iovita[1], Chen Feng[1,✉], Jing Zhang[1,✉]

[1]New York University [2]Yale University [3]Max Planck Institute for Evolutionary Anthropology

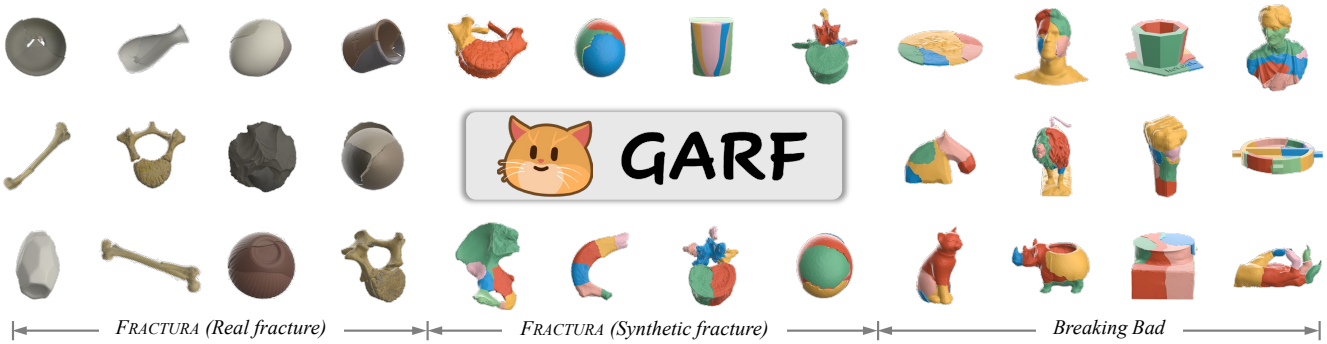

Figure 1. We curate FRACTURA, a unique dataset presenting real-world fracture assembly challenges across scientific domains, including ceramics, bones, eggshells, and lithics. To tackle these challenges, we introduce GARF, a *generalizable* 3D reassembly framework designed to handle varying *object shapes*, diverse *fracture types*, and the presence of *missing* or *extraneous* fragments.

Abstract

3D reassembly is a challenging spatial intelligence task with broad applications across scientific domains. While large-scale synthetic datasets have fueled promising learning-based approaches, their generalizability to different domains is limited. Critically, it remains uncertain whether models trained on synthetic datasets can generalize to real-world fractures where breakage patterns are more complex. To bridge this gap, we propose GARF*, a* ***g****eneralizable 3D re****a****ssembly framework for* ***r****eal-world* ***f****ractures.* GARF *leverages fracture-aware pretraining to learn fracture features from individual fragments, with flow matching enabling precise 6-DoF alignments. At inference time, we introduce one-step preassembly, improving robustness to unseen objects and varying numbers of fractures. In collaboration with archaeologists, paleoanthropologists, and ornithologists, we curate* FRACTURA*, a diverse dataset for vision and learning communities, featuring real-world fracture types across ceramics, bones, eggshells, and lithics. Comprehensive experiments have shown our approach consistently outperforms state-of-the-art methods on both synthetic and real-world datasets, achieving 82.87% lower rotation error and 25.15% higher part accuracy. This sheds light on training on synthetic data to advance real-world 3D puzzle solving, demonstrating its strong generalization across unseen object shapes and diverse fracture types.* GARF*'s videos, demo and code are available here.*

1. Introduction

We have long been captivated by questions of human origins [52]: *What are we? Where do we come from?* The answers to these fundamental questions lie in archaeological materials such as bones [6], ceramics [50], and lithics [34]. However, these artifacts are often highly fragmented and incomplete [6, 19]. Reassembling them requires placing each fragment in its correct *position* and *orientation* to restore a complete or functional entity [41]. This process is not only time-consuming but also challenges the limits of human *spatial intelligence* [10, 41, 47]. For instance, an experimental study on lithic refitting reported a success rate of only 30%, with experts performing only marginally better than novices [20]. Consequently, museum storerooms around the world remain filled with thousands of unassembled fragments, waiting to be pieced back together [47].

Recently, the emergence of the large-scale dataset Breaking Bad [42] has brought new hope to this domain, fueling the development of data-intensive reassembly methods [7, 17, 21, 23, 28, 30, 41, 49, 51, 54, 57, 61]. While PuzzleFusion++ achieves state-of-the-art (SOTA) performance

*, † Equal contribution.
✉ Corresponding authors {z.jing, cfeng}@nyu.edu.

on the *everyday* subset, its generalization degrades significantly on the *artifact* subset due to its reliance on global geometry learning [51]. More critically, the Breaking Bad dataset is generated via physics-based simulation [43], raising a fundamental question: *can models trained on synthetic breakage patterns generalize to real-world fractures*?

To answer this question, we identify two major real-world fracture challenges and curate FRACTURA, a dataset capturing key complexities: **(i) Data diversity** encompasses three geometrically distinct fracture types across multiple scientific domains, including bones, ceramics, eggshells, and lithics. As shown in Fig. 2, ceramics exhibit irregular, chaotic fractures typical of random breakage events, whereas flintknapping produces conchoidal fractures. This allows a systematic study of how *object shapes and fracture types* affect reassembly performance. **(ii) Missing or extraneous fragments** are common issues in real-world scenarios [13, 47] (see Fig. 2), which guides our model design to improve robustness.

In response to these complexities, we propose GARF, a **g**eneralizable 3D re**a**ssembly framework for **r**eal-world **f**ractures, featuring four components: **(i) Large-scale fracture-aware pretraining** takes lessons from recent successes of large-scale pretraining in natural language and computer vision [1, 35, 40, 44, 45]. This module learns fracture segmentation of individual fragments to handle unseen objects as well as missing or extraneous fragments. **(ii) Flow-based reassembly on** $\mathrm{SE}(3)$ introduces flow matching to learn pose distribution, leveraging the $\mathrm{SO}(3)$ manifold for accurate rotation estimation. Inspired by human puzzle-solving, we design a multi-anchor training strategy that randomly selects a subset of fragments to form local structures, exposing the model to diverse combinations to enhance distribution learning. **(iii) One-step pre-assembly at inference time** emphasizes the importance of pose initialization. Unlike image generation methods using a verifier to select optimal noise, we observe that the one-step output of flow matching provides a strong pose initialization. **(iv) LoRA-based fine-tuning** enhances the model's adaptability to specific domains.

Our main contributions are as follows:

- We introduce GARF, the first flow-based 3D fracture assembly framework that integrates fracture-aware pretraining and test-time one-step pre-assembly, achieving SOTA performance across synthetic and real-world datasets. GARF reduces rotation error by 82.87% and improves part accuracy by 25.15% compared to previous methods.
- GARF sheds light on training on synthetic data for real-world challenges, effectively handling unseen object shapes and diverse fracture types, while remaining robust to missing or extraneous fragments.
- Collaborating with domain experts, we curate FRACTURA, a diverse dataset capturing real-world fracture complexities, to conduct the first study on 3D reassembly generalization across ceramics, bones, eggshells, and lithics. Moreover, we apply the first integration of LoRA-based fine-tuning for domain-adaptive 3D reassembly.

Table 1. **Comparisons of 3D Reassembly Datasets.** For Fracture Type*: Syn. denotes synthetic data. The number in parentheses indicates the number of synthetic/real fracture modes.

Datasets	Breaking Bad [42]	Fantastic Breaks [18]	RePAIR [47]	FRACTURA
# Pieces	8M	300	1070	53350+292
# Breaks	2-100	2	2-44	2-22
# Assemblies	10474	150	117	9727+41
Fracture Type*	Syn. (1)	Real (1)	Real (1)	**Syn. (2) + Real (3)**
Object Type	Everyday Artifact, Other	Everyday	Frescoes	Bones, Eggshells Lithics, Ceramics
Miss. / Extra.	× / ×	× / ×	✓ / ×	✓ / ✓
Texture	×	×	✓	✓

2. FRACTURA Dataset

Existing datasets, whether synthetic [42] or real [18, 47], are limited to a single fracture type. To fill this gap, we curate FRACTURA, a challenging dataset (see Fig. 2), designed for a comprehensive evaluation of how *object shapes*, *fracture types*, and the presence of *missing* or *extraneous* parts affect reassembly performance.

Data Characteristics. FRACTURA comprises both real and synthetic fracture subsets. In collaboration with archaeologists, paleoanthropologists, and ornithologists, the **real fracture** subset includes three real fracture types relevant to scientific challenges (see Fig. 2): (i) *Random breakage* produces irregular, chaotic fractures, commonly observed in *ceramics*, *bones*, and *eggshells*. (ii) *Incomplete ossification* results in unfused bone ends (epiphyses) in *juvenile skeletons*, leading to fragmented rather than intact bones in skeletal collections and fossil records. Reassembling these unfused parts will benefit the following analysis and further create a complete series of bone developments over time from early childhood to adulthood. (iii) *Flintnapping* produces conchoidal fractures in *lithics*, characterized by radially propagating fracture lines (see Fig. 2). Real-world collections and scans naturally incorporate *missing or extraneous fragments*. For the **synthetic fracture** subset, we generate realistic fractures using physics-based simulation [43] for *ceramics*, *bones*, and *eggshells*, and geometry-based simulation [36] for *lithics*.

Data Collection and Simulation. We utilize the high-accuracy Artec spider 3D scanner (precision: 0.05 mm) to acquire detailed 3D meshes of real fragments and intact objects from the same categories. The real fracture subset serves as test data, while intact objects are used to generate the synthetic fracture subset for fine-tuning and evaluation. Details are provided in the supplementary materials.

Data Statistics. Table 1 summarizes key statistics of FRACTURA and other 3D reassembly datasets. Compared with existing datasets, FRACTURA introduces a diverse set of

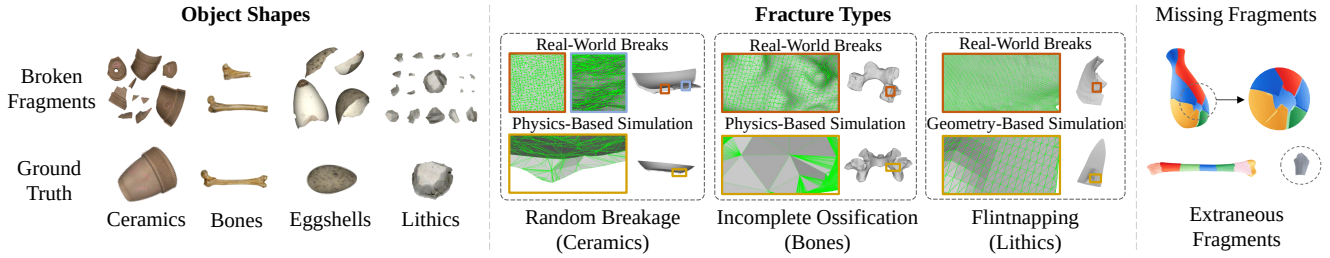

Figure 2. **Characteristics and Challenges of FRACTURA.** (i) Diverse fracture types including two synthetic and three real-world types, across ceramics, bones, eggshells, and lithics. (ii) Real-world challenges such as missing or extraneous fragments.

real and synthetic fractures across multiple scientific domains. We are actively expanding its size and diversity. More details are provided in the supplementary materials.

3. Method

Previous work either enhances global geometry learning through *fragment features* [41, 51, 54] or *jointly* learns hierarchical features from both global and local geometry [30]. In contrast, as shown in Fig. 3, GARF decouples the understanding of *local fracture* features (Sec. 3.1) and *global* fragment alignment (Sec. 3.2). At inference, we propose a one-step pre-assembly strategy (Sec. 3.3) for robustness to unseen objects and increasing numbers of fractures. To further boost the performance on domain-specific data, we employ a LoRA-based fine-tuning method (Sec. 3.4).

3.1. Why Large-Scale Fracture-Aware Pretraining?

Humans can infer fracture points without knowledge of the global object shape [30, 37]. To emulate this ability, we leverage large-scale data to learn local fracture features from individual fragments, drawing inspiration from recent advances in large-scale pretraining [1, 35, 40, 44, 45]. This enables our method to generalize to unseen object shapes and handle missing or extraneous fragments.

Specifically, we sample a set of point clouds $\boldsymbol{P} = \{P^1, P^2, \ldots, P^N\}$ to represent fragments, where N is the number of fragments. To extract fracture-level features from $\boldsymbol{P}$, we employ Point Transformer V3 (PTv3) [55] as the backbone, integrating two MLP layers [30] as the segmentation head for fracture cloud segmentation. Since fracture points constitute only a small proportion of $\boldsymbol{P}$, an imbalance arises between positive and negative samples. To mitigate this, we adopt the Dice loss function [46]:

$$\mathcal{L}_{\text{Seg}} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i + \epsilon}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i + \epsilon}, \tag{1}$$

where p is the predicted value and g is its ground truth label. To derive the high-quality g, we directly extract shared surfaces between connected fragments from the mesh, defining them as fracture surfaces $\boldsymbol{F}$. We then apply Poisson disk sampling [5] to generate $\boldsymbol{P}$, ensuring that $g = \boldsymbol{P} \cap \boldsymbol{F}$. This uniform sampling prevents the encoder from overfitting to specific point densities, improving generalization.

3.2. Flow-Based Reassembly on SE(3)

Inspired by flow-based generative models [25–27, 38] in image synthesis [9], protein structure generation [4, 11, 16], and robotic manipulation [3, 58], we leverage flow matching (FM) to model the fracture reassembly process.

On a manifold $\mathcal{M}$, the flow $\psi_t : \mathcal{M} \to \mathcal{M}$ is defined as the solution of an ordinary differential equation (ODE):

$$\frac{\mathrm{d}}{\mathrm{d}t}\psi_t(\boldsymbol{x}) = \boldsymbol{v}_t(\psi_t(\boldsymbol{x})), \quad \psi_0(\boldsymbol{x}) = \boldsymbol{x}, \tag{2}$$

where $\boldsymbol{v}_t(\boldsymbol{x}) \in \mathcal{T}_{\boldsymbol{x}}\mathcal{M}$ is the time-dependent vector field, and $\mathcal{T}_{\boldsymbol{x}}\mathcal{M}$ denotes the tangent space at $\boldsymbol{x}$. In the context of $\mathrm{SE}(3)$, the tangent space corresponds to the *Lie algebra* $\mathfrak{se}(3)$, a six-dimensional vector space representing the velocity of the rigid motion of fragments.

Given an object composed of N fragments, we represent their poses as $\boldsymbol{T} = \{T^1, T^2, \cdots, T^N\}$, where each T^i consists of a rotation $r \in \mathrm{SO}(3)$ and a translation $a \in \mathbb{R}^3$, expressed as $T^i : \{r, a\} \in \mathrm{SE}(3)$. The initial noise distribution is defined as: $p_0(\boldsymbol{T}_0) = \mathcal{U}(\mathrm{SO}(3)) \otimes \mathcal{N}(0, I_3)$, where the rotation noise follows a uniform distribution over $\mathrm{SO}(3)$, and the translation noise is sampled from a unit Gaussian distribution. We decouple the rotation and translation flows, allowing independent flow modeling in $\mathrm{SO}(3)$ and $\mathbb{R}^3$. Therefore, the conditional flow $\boldsymbol{T}_t = \psi_t(\boldsymbol{T}_0 \mid \boldsymbol{T}_1)$ follows the geodesic path connecting $\boldsymbol{T}_0$ and $\boldsymbol{T}_1$:

$$\begin{aligned} \boldsymbol{r}_t &= \exp_{\boldsymbol{r}_0}(t \log_{\boldsymbol{r}_0}(\boldsymbol{r}_1)), \\ \boldsymbol{a}_t &= (1-t)\boldsymbol{a}_0 + t\boldsymbol{a}_1, \end{aligned} \tag{3}$$

where $\exp_{\boldsymbol{r}}$ and $\log_{\boldsymbol{r}}$ are the exponential and logarithmic maps on $\mathrm{SO}(3)$. The final optimization objective is:

$$\begin{aligned} \mathcal{L}_{\text{FM}} = &\mathbb{E}_{t, p_1(\boldsymbol{T}_1), p_t(\boldsymbol{T}|\boldsymbol{T}_1)} \left[\sum_{i=1}^{N} \left\| \boldsymbol{v}_r^i(\boldsymbol{T}_t, t) - \frac{\log_{\boldsymbol{r}_t}(\boldsymbol{r}_1)}{1-t} \right\|_g^2 \right. \\ &\left. + \left\| \boldsymbol{v}_a^i(\boldsymbol{T}_t, t) - \frac{\boldsymbol{a}_1 - \boldsymbol{a}_t}{1-t} \right\|_g^2 \right]. \end{aligned} \tag{4}$$

Further details are in the supplementary materials.

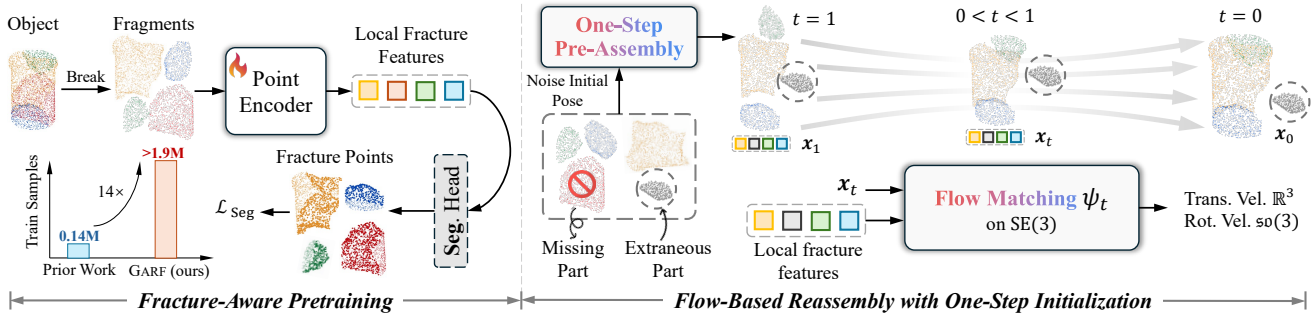

Figure 3. **Pipeline of GARF.** Our framework comprises two main components: (i) Fracture-aware pretraining leverages 14× more data than previous methods to learn the local fracture features via fracture point segmentation, and (ii) Flow-based reassembly on SE(3) leverages the SO(3) manifold for precise rotation estimation. At inference time, one-step pre-assembly strategy provides better initial poses, enhancing robustness against unseen objects and increasing numbers of fractures.

Network Architecture. Consider an object consisting of N fragments with inital poses $\boldsymbol{T}_t \in \mathbb{R}^{N\times 7}$ at timesteps $\boldsymbol{t}$. The corresponding latent features are extracted from the pre-trained encoder: $\boldsymbol{F} = \mathcal{E}(\boldsymbol{P}) \in \mathbb{R}^{M\times c}$, where c is the number of channels and M is the pre-defined number of sampled points. We integrate point cloud coordinates $\boldsymbol{P} \in \mathbb{R}^{M\times 3}$, normals $\boldsymbol{n} \in \mathbb{R}^{M\times 3}$, and scale information $\boldsymbol{s} \in \mathbb{R}^{M\times 1}$ as pose-invariant shape priors in the position embedding:

$$\boldsymbol{s}_{\text{emb}} = f_{\text{shape}}\Big(\text{concat}\Big(\boldsymbol{F},\, \texttt{PE}(\boldsymbol{P}),\, \texttt{PE}(\boldsymbol{n}),\, \texttt{PE}(\boldsymbol{s})\Big)\Big). \tag{5}$$

The pose is treated as spatial information: $\boldsymbol{p}_{\text{emb}} = f_{\text{pose}}\big(\texttt{PE}(\boldsymbol{T})\big)$. These embeddings are combined to form the Transformer input: $\boldsymbol{d} = \texttt{PE}(\boldsymbol{s}_{\text{emb}} + \boldsymbol{p}_{\text{emb}})$.

The feature $\boldsymbol{d}$ is then processed through L Transformer layers including self-attention and global attention. Self-attention adopts efficient execution of FlashAttention [8]:

$$\begin{aligned} \boldsymbol{Q},\, \boldsymbol{K},\, \boldsymbol{V} &= \text{Linear}(\text{LN}(\boldsymbol{d})), \\ \boldsymbol{A}_{\text{self}} &= \text{FlashAttn}\Big(\boldsymbol{Q},\, \boldsymbol{K},\, \boldsymbol{V};\, \text{cu}(\ell),\, \ell_{\max}\Big), \end{aligned} \tag{6}$$

where the model computes the sequence lengths ℓ of the variable number of sampled points per fragment batch, along with their cumulative sums $\text{cu}(\ell)$. The self-attention output $\boldsymbol{A}_{\text{self}}$ is added to the original feature $\boldsymbol{d}$, yielding the updated representation $\boldsymbol{h} = \boldsymbol{d} + \boldsymbol{A}_{\text{self}}$. The global attention layer then applies a similar attention mechanism to aggregate information across fragments with $\ell = M$. The output undergoes Layer Normalization (LN) and a feed-forward network (FFN) with a residual connection, before regressing the pose $\boldsymbol{T}_{t-1} : \{\boldsymbol{r}_{t-1}, \boldsymbol{a}_{t-1}\}$ at the next timestep:

$$\begin{aligned} \boldsymbol{h} &\leftarrow \boldsymbol{h} + \text{FFN}(\text{LN}(\boldsymbol{h})), \\ \boldsymbol{a}_{t-1} &= f_{\text{trans}}(\boldsymbol{h}), \quad \boldsymbol{r}_{t-1} = f_{\text{rot}}(\boldsymbol{h}). \end{aligned} \tag{7}$$

Multi-Anchor Training Strategy. We observe that probability paths vary significantly across fragments; some require complex transformations, while others remain nearly stationary. To model this variation, we introduce a multi-anchor training strategy, randomly selecting $k \in [1, N-1]$ fragments as anchors and fixing their positions with identity rotations and zero translations. For these anchor fragments $\boldsymbol{i}$, the vector field is explicitly supervised to be zero: $\boldsymbol{v}^{\boldsymbol{i}}(\boldsymbol{T}_t, t) = 0$. Unlike prior approaches [51], which prevent gradient propagation for a single anchor (the largest fragment and its connected neighbors within a 50% threshold), our multi-anchor strategy enforces a zero vector field, expanding the range of probability paths and enhancing generalization across diverse fragment configurations.

3.3. One-Step Pre-Assembly at Inference Time

In diffusion-based and flow-based models, prior work in image generation emphasizes the role of inference-time scaling, employing search strategies and self-supervised verifiers to find better noise candidates [24, 32, 39, 62]. In contrast, we propose a simple yet effective one-step pre-assembly at inference time, leveraging FM for its ability to model straight-line probability paths. Specifically, we perform a one-step FM inference (step $= 1$) to generate an initial pose $\boldsymbol{T}_0'$, followed by the standard multi-step flow for refinement. This pre-assembly step effectively narrows the search space, improving pose initialization for subsequent refinement. Despite a minimal 5% increase in computational cost ($20 \rightarrow 1 + 20$ steps), our experiments show that this strategy significantly enhances assembly performance, particularly for larger fragment sets.

3.4. LoRA-based Fine-tuning

To quickly adapt to domain-specific contexts, we employ a LoRA-based [15] fine-tuning approach using the synthetic fracture subset in FRACTURA. Specifically, we integrate LoRA adapters into the self-attention and global attention layers of the final Transformer block while unfreezing the MLP heads for pose prediction (f_{rot} and f_{trans}). Our experiments demonstrate that this lightweight fine-tuning method requires as few as 5–10 domain-specific objects to achieve

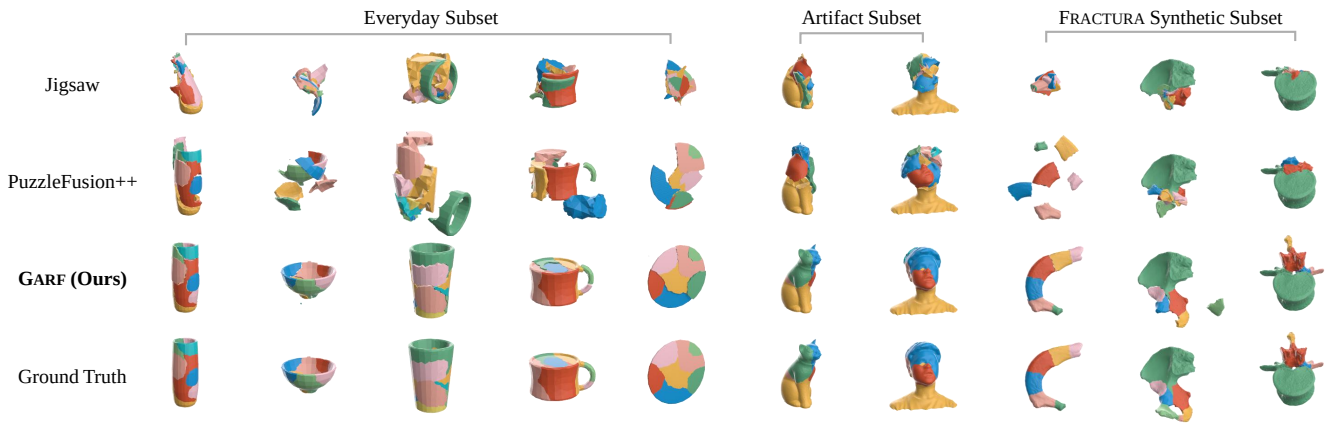

Figure 4. **Qualitative Comparisons on the Breaking Bad and FRACTURA.** GARF consistently produces more accurate reassemblies, particularly on the Breaking Bad Artifact subset and FRACTURA synthetic fracture subset, demonstrating strong generalization to unseen object shapes. Meshes are used for visualization only. Additional results are available in the supplementary material.

substantial improvements in scientific applications such as juvenile skeleton reconstruction and lithic refitting.

4. Experiments

4.1. Training and Evaluation Details.

Training. For fracture-aware pretraining, Point Transformer V3 (PTv3) [55] serves as the backbone, extracting 64-channel features from its final layer. GARF is pretrained on three Breaking Bad [42] subsets, totaling 1.9M fragments—14× more than prior works. For fair comparisons, GARF-mini is pretrained only on the Everyday subset. For FM, both GARF and GARF-mini are trained on the Everyday subset. We use a standard transformer [51] to compute the vector field, with each block consisting of 6 encoder layers, 8 attention heads per layer, and an embedding dimension of 512. The initial learning rate is 2e-4 and decays by a factor of 2 at epochs 900 and 1200. GARF is trained with a batch size of 32 on 4 NVIDIA H100 HBM3 GPUs, requiring 2 days for pretraining and 3 days for FM. GARF-mini completes pretraining in 0.5 days. For LoRA fine-tuning, we use the PEFT framework [33] with rank $r = 128$, $\alpha = 256$, and a dropout rate of 0.1.

Datasets. We evaluate our model on three datasets with diverse object shapes and fracture types: **(i) Breaking Bad** [42], the largest synthetic fracture dataset for 3D reassembly. We use the volume-constrained version, evaluating 7,872 assemblies from the Everyday subset and 3,697 assemblies from the Artifact subset. Results on the vanilla version are in the supplementary materials. **(ii) Fantastic Breaks** [18], a real-world dataset of 195 manually scanned fractured objects with complex surfaces, used for evaluation only. **(iii)** FRACTURA, a mixed synthetic-real dataset spanning ceramics, bones (vertebrae, limbs, ribs), eggshells, and lithics. For the synthetic fracture subset, we follow an 80/20 split [47] for LoRA fine-tuning and evaluation.

Evaluation Metrics. Following [51], we evaluate assembly quality using four metrics: (i) RMSE(R) is the root mean square error of rotation (degrees); (ii) RMSE(T) is the root mean square error of translation; (iii) PA is the percentage of correctly assembled fragments, where the per-fragment chamfer distance is below 0.01; and (iv) CD is the chamfer distance between the assembled object and ground truth.

Competing Methods. We compare our approach against Global [22], LSTM [53], DGL [60], Jigsaw [30], PMTR* [21], and PuzzleFusion++ (PF++) [51]. We implemented PF++ [51] and Jigsaw [30] using its official codebase, while performance metrics for other methods on the volume-constrained Breaking Bad dataset are sourced from their official papers or repositories. Additional comparisons with SE(3)-Equiv [54], DiffAssemble [41], and PH-Former [7] on the vanilla Breaking Bad dataset, as well as FragmentDiff [57] on its custom Breaking Bad subset, are provided in the supplementary materials.

4.2. Can GARF Generalize to Unseen Shapes?

Real-world fragmentation varies in object geometries, requiring GARF to *generalize to unseen object shapes*. To empirically assess this, we evaluate GARF-mini on FRACTURA and the Artifact subset of Breaking Bad, as well as GARF on FRACTURA.

Results and Analysis. Table 2 presents the evaluation results on the Breaking Bad dataset, where GARF achieves significant improvements over previous SOTA methods. Compared to PF++ [51], GARF reduces rotation error by 82.87% and translation error by 79.83%, while achieving a CD below 0.001, indicating near-perfect reconstructions. Even more impressively, GARF-mini demonstrates exceptional generalization capability. Despite being trained only on the Everyday subset, it maintains consistent per-

*The performance of PMTR [21] is from PuzzleFusion++ [51].

Table 2. **Quantitative Results on Volume-Constrained Breaking Bad [42] and FRACTURA Datasets.** The best performance metric is highlighted in **bold**, while the second-best is underlined.

Methods	RMSE(R) ↓ degree	RMSE(T) ↓ $\times 10^{-2}$	PA ↑ %	CD ↓ $\times 10^{-3}$
Tested on the **Everyday** Subset				
Global [22]	80.50	14.60	28.70	13.00
LSTM [53]	82.70	15.10	27.50	13.30
DGL [60]	80.30	13.90	31.60	11.80
Jigsaw [30]	42.19	6.85	68.89	8.22
PMTR [21]	31.57	9.95	70.60	5.56
PF++ [51]	35.61	6.05	76.17	2.78
GARF-mini	6.68	1.34	94.77	0.25
GARF	**6.10**	**1.22**	**95.33**	**0.22**
Tested on the **Artifact** Subset				
Jigsaw	43.75	7.91	65.12	8.50
PF++	47.03	10.63	57.97	8.24
GARF-mini	7.67	1.77	93.34	0.81
GARF	**5.82**	**1.27**	**95.04**	**0.42**
Tested on the **Fractura** (Synthetic Fracture)				
Jigsaw	60.50	18.49	33.06	70.68
PF++	62.57	18.65	37.74	36.13
GARF-mini	27.88	6.79	76.25	7.70
GARF	**19.63**	**4.93**	**83.41**	**6.06**

formance on the *unseen* Artifact subset, avoiding the performance degradation typically observed with unseen object shapes. This highlights GARF's robust feature extraction and assembly mechanisms. Notably, GARF achieves 95.33% and 95.04% PA on the Breaking Bad dataset, approaching the theoretical maximum PA of 96.49% (Everyday subset) and 96.10% (Artifact subset) when excluding fragments smaller than 0.1% of the object volume[†]. On the challenging synthetic subset of FRACTURA, which contains *unseen domain-specific objects*, GARF further demonstrates superior generalization capability, outperforming all competing methods across all evaluation metrics. As shown in Fig. 4, GARF consistently produces more accurate reassemblies, particularly on the Breaking Bad Artifact subset and the FRACTURA synthetic fracture subset, confirming its strong generalization to unseen object shapes.

4.3. How Do Fracture Types Affect Generalization?

As scanning time significantly increases with fragment count [18], real fracture datasets are typically limited in *size* and *diversity*, making large-scale training infeasible. To address this, we investigate *how fracture types impact zero-shot generalization* on Fantastic Breaks and FRACTURA, as well as *the fine-tuning performance using domain-specific synthetic fractures* on FRACTURA.

Real Fracture on Fantastic Breaks [18]. As shown in Table 3, GARF demonstrates superior generalization to real-world fracture surfaces, achieving a remarkable 48.65% reduction in rotation error compared to PF++ [51], indicating that our model effectively bridges the synthetic-to-real gap in fracture surface understanding for everyday objects.

[†]Scanning tiny 3D fragments in real-world settings is inherently challenging; anthropologists often treat such fragments as missing parts.

Table 3. **Quantitative Results on Fantastic Breaks Dataset [18].** This includes manually collected real-world objects.

Methods	RMSE(R) ↓ degree	RMSE(T) ↓ $\times 10^{-2}$	PA ↑ %	CD ↓ $\times 10^{-3}$
Tested on the **Fantastic Breaks**				
Jigsaw [30]	26.30	6.43	73.64	10.47
PF++ [51]	20.68	4.37	83.33	6.68
GARF	**10.62**	**2.10**	**91.00**	**2.12**

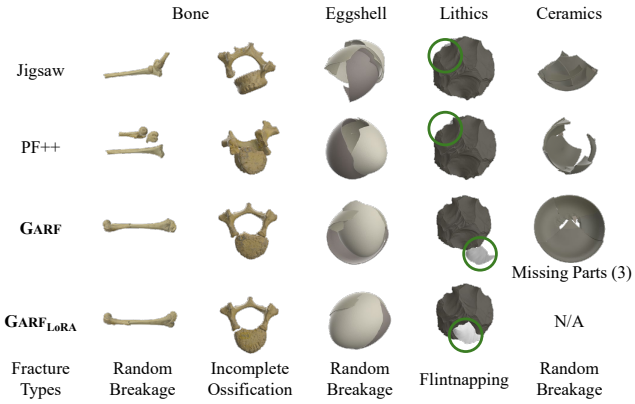

Figure 5. **Qualitative Comparisons on the FRACTURA real fracture subset.** GARF generalizes well to random breakage (limb bones and ceramics) and incomplete ossification (vertebrae) but faces challenges with high-ambiguity fractures like flintknapping (lithics). Fine-tuning enhances performance, particularly for thin-shell structures (eggshells) and flintnapping (lithics).

Real Fracture on FRACTURA. We further evaluate performance across three fracture types. To isolate the impact of unseen objects, GARF$_{LoRA}$ is fine-tuned separately on synthetic fractures from bones, eggshells, and lithics in FRACTURA[‡]. Figure 5 compares reassembly performance across three fracture types in the FRACTURA real fracture subset. GARF outperforms competing methods on ceramics even with three missing fragments. GARF$_{LoRA}$ further improves performance by mitigating the effect of unseen objects. GARF$_{LoRA}$ generalize well to random breakage (limb bones, ceramics) and incomplete ossification (vertebrae). Unfortunately, although finetune significantly improves the performance on lithics, both GARF and GARF$_{LoRA}$ struggle on lithics, likely due to high ambiguity among flakes and the core, a well-known challenging spatial reasoning problem for anthropologists [20]. This unresolved challenge in FRACTURA offer valuable directions for future research.

[‡]Fine-tuning is not performed for ceramics, as its object categories closely resemble those in the Everyday subset of Breaking Bad.

Table 4. **Quantitative Results on Missing / Extraneous Parts.**

Methods	Input	RMSE(R) ↓	RMSE(T) ↓	PA ↑	CD ↓
Jigsaw [30]	Complete	71.41	15.83	28.34	21.94
	20% Miss.	70.45	15.54	28.85	21.58
	20% Extra.	74.55	19.02	24.03	26.15
PF++ [51]	Complete	59.26	12.00	49.38	5.52
	20% Miss.	61.14	12.26	44.71	7.57
	20% Extra.	61.25	13.91	40.59	8.77
GARF	Complete	19.55	3.83	83.39	0.62
	20% Miss.	**22.23**	**4.75**	**78.87**	**1.40**
	20% Extra.	**22.70**	**4.62**	**79.21**	**1.29**

4.4. How Do Missing or Extraneous Parts Affect Performance?

Archaeological materials are often *incomplete* or mixed with similar but *extraneous* fragments, posing significant challenges for assembly. To quantitatively assess model robustness under these conditions at scale, we extend Breaking Bad's Everyday subset in two ways: (i) *Missing* parts subset removes 20% of fragments in descending order of volume, preserving the largest anchor fragments and maintaining the object's connectivity graph; (ii) *Extraneous* parts subset adds fragments from other objects in the same category, selecting pieces smaller than the largest anchor fragment but larger than 5% of the object's total volume to ensure they are not trivially small. For a fair comparison, we evaluate objects with 5 to 17 fragments. Additionally, we provide visual demonstrations of assemblies with naturally missing or extraneous parts in FRACTURA.

Results and Analysis. Table 4 shows that GARF demonstrates strong resilience, with minimal performance degradation over competing methods. With 20% extraneous fragments, GARF maintains a high PA of 79.21%, only 5.0% lower than with complete sets, whereas PF++ drops by 24.32%. Similarly, with 20% missing fragments, GARF achieves 78.87% PA, far surpassing Jigsaw (28.85%) and PF++ (47.22%). Figure 6 illustrates how missing or extraneous fragments affect reassembly performance. While all methods degrade under these challenging conditions, GARF demonstrates superior robustness, maintaining coherent structures despite missing or extraneous fragments. In contrast, Jigsaw and PF++ exhibit severe misalignments and fragment mismatches. This robustness suggests that GARF can partially handle missing or extraneous fragments, benefiting from our model design.

4.5. Ablation Study

Pretraining Strategy. We evaluate our fracture-aware pretraining strategy by comparing PF++ [51] with two variants using the same diffusion model: (i) VAE-based pretraining [48], and (ii) our fracture-aware pretraining. As shown in Table 5, our strategy reduces RMSE(R) by 53.33% and RMSE(T) by 49.05% compared to the VAE-based strategy.

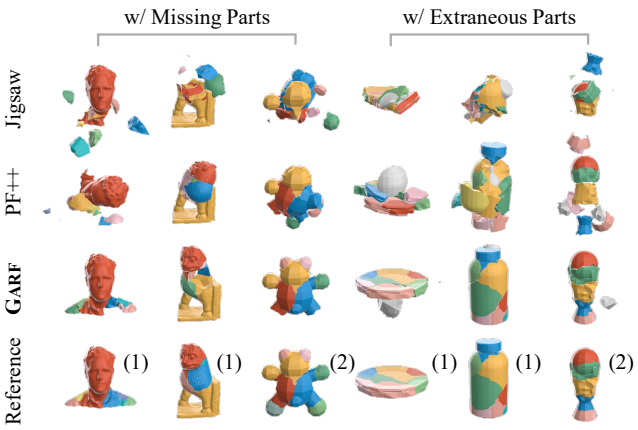

Figure 6. **Qualitative Comparisons on the Missing or Extraneous Impact.** GARF demonstrates superior robustness, maintaining coherent structures despite missing or extraneous fragments.

Table 5. **Ablation Study on the Fracture-Aware Pretraining.**

	RMSE(R) ↓	RMSE(T) ↓	PA ↑	CD ↓
PF++ (VAE)	32.91	5.26	78.95	3.04
PF++ (**Fracture-aware**)	**15.36**	**2.68**	**89.40**	**0.66**

Table 6. **Ablation Study on Our Designs of FM.**

SE(3)	Multi-Anchor	One-Step	RMSE(R) ↓	RMSE(T) ↓	PA ↑
✗	✗	✗	10.24	1.95	89.08
✓	✗	✗	8.02	1.63	93.78
✓	✓	✗	7.63	1.60	94.02
✓	✓	✓	**6.68**	**1.34**	**94.77**

Designs in FM. We conduct an ablation study to evaluate the impact of design choices in our FM module. As shown in Table 6, vanilla FM, trained with spherical linear interpolation (slerp) to approximate valid rotations in the forward process [12], achieves 89.08 PA, already surpassing previous methods [30, 51]. Incorporating the SE(3) representation further improves performance by pre-modeling the manifold distribution and better capturing distribution shifts during assembly. Multi-anchor training strategy further enhances results, while one-step pre-assembly significantly boosts performance by providing a more reasonable initial pose distribution, leading to the best overall outcomes.

Sample Steps / One-Step Initialization. Table 7 shows the effect of varying sampling steps in our FM framework. Surprisingly, even with just 5 steps, FM achieves 93.70% PA, highlighting its effectiveness in modeling global probabilistic paths. Additionally, our one-step pre-assembly provides a more reasonable initial pose, further improving assembly quality while adding minimal computational overhead.

Anchor Fragment. Similar to PF++ [51], we use the largest fragment as the anchor fragment at inference. We compare the performance of using the largest fragment, a randomly selected fragment, and no anchor fragment. As shown in Table 8, using a random fragment as the anchor fragment has almost no negative effect on the model. Only

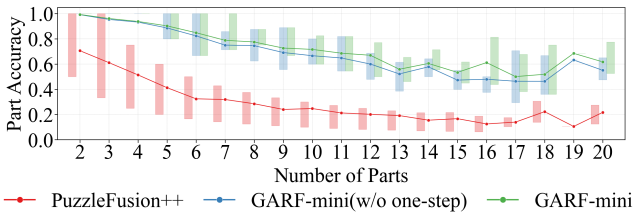

Figure 7. **Performance vs. Fragment Count on the Artifact Subset.** GARF-mini significantly outperforms PF++, with one-step pre-assembly further boosting results for > 10 fragments.

anchor-free initialization leads to a slight performance drop.

Comparison with Diffusion Models. Table 9 compares our FM module with diffusion models. While diffusion, when paired with fracture-aware pretraining, achieves competitive performance, directly applying vanilla FM yields lower results (89.08 PA), emphasizing the importance of our subsequent design choices. A key limitation of diffusion models is their handling of $\mathrm{SO}(3)$ rotation, which cannot be naturally incorporated into the reverse process. Existing methods, such as score prediction [59], aim to maintain rotation validity but fall outside our current scope. Additionally, diffusion models rely on multi-step denoising without explicitly modeling the global probabilistic path, rendering one-step pre-assembly ineffective. Furthermore, on FRACTURA, diffusion models exhibit weaker generalization to unseen objects compared to GARF-mini.

Number of Fragments. We analyze performance across varying fragment counts, as shown in Figure 7. GARF-mini consistently surpasses PF++ across all fragment counts. Our one-step pre-assembly further enhances performance on unseen objects containing more than 10 fragments.

5. Impact and Limitations

Scientific Impacts. How objects—whether they are bones, ceramic pots, or stone tools—were reassembled and what processes influenced their reconstruction is one of the basic questions common to different research communities, including *paleontology* and *paleoanthropology*, *archaeology*, and *forensic science*. To explore this, we make the first attempt to collaborate with archaeologists, paleoanthropologists, and ornithologists to build a generalizable model for real-world fracture reassembly. While GARF achieves significant improvements, challenges remain, particularly in handling the complexities presented by FRACTURA, which creates new opportunities for the vision and learning communities, encouraging advancements in 3D puzzle solving.

Limitations and Future Directions. GARF encounters challenges in handling fracture ambiguity, especially when fragments have subtle geometric differences or inherently ambiguous fracture surfaces. For instance, it struggles with lithic refitting in FRACTURA due to high ambiguity among flakes and the core, as well as fresco reconstruction in REPAIR, where erosion affects fracture surfaces [47]. Therefore, our future work will focus on: (i) Multimodal fracture reassembly, integrating geometric and texture information; (ii) Test-time policy optimization using expert feedback; (iii) Expanding the size and diversity of FRACTURA.

Table 7. **Ablation Study on Sample Steps.**

Steps	RMSE(R) ↓	RMSE(T) ↓	PA ↑	CD ↓	Speed (ms)
1	12.52	3.18	86.88	2.14	38.26
One-Step + 1	9.79	2.46	91.31	1.42	45.76
5	8.25	1.92	93.70	0.53	57.32
One-Step + 5	7.15	1.66	94.43	0.46	76.23
20	7.63	1.60	94.02	0.35	185.05
One-step + 20	**6.68**	**1.34**	**94.77**	**0.25**	190.77
50	7.50	1.54	94.01	0.32	408.40

Table 8. **Ablation Study on the Different Anchor Initialization.**

Settings	RMSE(R) ↓	RMSE(T) ↓	PA ↑	CD ↓
Largest Anchor	6.10	1.22	95.33	0.22
Random Anchor	6.09	1.30	95.20	0.29
Anchor-Free	9.09	2.13	93.23	0.91

Table 9. **Comparison Between Diffusion and Our FM Models.**

Dataset	Methods	RMSE(R) ↓	RMSE(T) ↓	PA ↑
Everyday	Diffusion	7.45	1.47	94.30
	SE(3) Diffusion	N/A	N/A	N/A
	Diffusion w/ One-Step	7.51	1.47	94.27
	Vanilla FM	10.24	1.95	89.08
	GARF-mini	**6.68**	**1.34**	**94.77**
FRACTURA	Diffusion	32.38	7.90	71.73
	GARF-mini	**27.88**	**6.79**	**76.25**

6. Related Work

Fracture Assembly. Fracture Assembly is a challenging spatial intelligence task. Early methods relied on explicit geometric matching with handcrafted features [2, 14, 31, 56], often struggling with ambiguous or incomplete geometries. The advent of the large-scale synthetic dataset Breaking Bad [42] has enabled learning-based approaches to acquire robust geometric representations and assembly strategies [29]. Jigsaw [30] jointly learns hierarchical features from global and local geometries for fracture matching and pose estimation, while SE(3)-equiv [54] extracts fragment features for pose estimation. DiffAssemble [41] improves performance using a diffusion model. PuzzleFusion++ [51] mimics how humans solve spatial puzzles by integrating diffusion-based pose estimation with a VAE-based fragment representation and transformer-based alignment verification. However, while PuzzleFusion++ achieves SOTA results on the everyday subset, its performance degrades significantly on the Artifact subset [51]. More critically, it remains unclear whether models trained on synthetic data can generalize to real-world fractures with more complex breakage patterns. To fill this gap, we identify major real-

world fracture challenges and curate FRACTURA, a dataset capturing key complexities. To address these challenges, we propose GARF, a generalizable 3D reassembly framework for real-world fractures.

Flow Matching. Flow matching (FM) has emerged as a powerful alternative to diffusion models, offering advantages such as simulation-free training, closed-form target vector fields, and more efficient optimization [25, 27, 38]. Unlike diffusion models that require iterative denoising, FM directly learns a vector field that smoothly transforms a prior distribution into the target distribution along straight probability paths. Recent advances have explored FM across image generation [9], protein backbone generation [11, 16, 58], and general robot control [3], demonstrating superior efficiency compared to diffusion-based alternatives. AssembleFlow [12] attempts to leverage FM for molecular assembly, but introduces numerical errors by approximating quaternion updates through direct addition over small time intervals during inference. This limitation has been analyzed in our ablation study (Section 4.5). While diffusion models have been widely applied to fracture assembly [41, 51], FM provides a more natural formulation by learning geodesic flows in SE(3). To leverage these advantages, we propose the first FM-based fracture assembly framework, incorporating a multi-anchor training strategy and a one-step pre-assembly at inference time.

7. Conclusion

In collaboration with archaeologists, paleoanthropologists, and ornithologists, we present FRACTURA, a diverse and challenging fracture assembly dataset, and GARF, a generalizable 3D reassembly framework designed for real-world fractures. FRACTURA serves as a challenging benchmark to evaluate how object shapes, fracture types, and the presence of missing or extraneous parts affect reassembly performance. Facing these challenges, GARF offers vital guidance on training on synthetic data to advance real-world 3D puzzle solving. Comprehensive evaluations demonstrate its strong generalization to unseen object shapes and diverse fracture types. Despite its superior performance, GARF still struggles with geometric ambiguity, particularly when dealing with highly similar fragments and eroded fracture surfaces. We anticipate that FRACTURA will drive further advancements in 3D reassembly, pushing the boundaries of spatial reasoning to answer unknown scientific questions.

Acknowledgement. We gratefully acknowledge the Physical Anthropology Unit, Universidad Complutense de Madrid for access to curated human skeletons, and Dr. Scott A. Williams (NYU Anthropology Department) for the processed data samples. This work was supported in part through NSF grants 2152565, 2238968, 2322242, and 2426993, and the NYU IT High Performance Computing resources, services, and staff expertise.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 3

[2] Anthousis Andreadis, Pavlos Mavridis, and Georgios Papaioannou. Facet extraction and classification for the reassembly of fractured 3d objects. In *Eurographics (Posters)*, pages 1–2, 2014. 8

[3] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. 3, 9

[4] Avishek Joey Bose, Tara Akhound-Sadegh, Guillaume Huguet, Kilian Fatras, Jarrid Rector-Brooks, Cheng-Hao Liu, Andrei Cristian Nica, Maksym Korablyov, Michael Bronstein, and Alexander Tong. Se (3)-stochastic flow matching for protein backbone generation. *arXiv preprint arXiv:2310.02391*, 2023. 3, 2

[5] Robert Bridson. Fast poisson disk sampling in arbitrary dimensions. *SIGGRAPH sketches*, 10(1):1, 2007. 3

[6] Ewen Callaway. Oldest homo sapiens fossil claim rewrites our species' history. *Nature*, 546:289–293, 2017. 1

[7] Wenting Cui, Runzhao Yao, and Shaoyi Du. Phformer: multi-fragment assembly using proxy-level hybrid transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1408–1416, 2024. 1, 5, 2

[8] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024. 4

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 3, 9

[10] Howard E Gardner. *Frames of mind: The theory of multiple intelligences*. Basic books, 2011. 1

[11] Tomas Geffner, Kieran Didi, Zuobai Zhang, Danny Reidenbach, Zhonglin Cao, Jason Yim, Mario Geiger, Christian Dallago, Emine Kucukbenli, Arash Vahdat, et al. Proteina: Scaling flow-based protein structure generative models. In *The Thirteenth International Conference on Learning Representations*. 3, 9, 2

[12] Hongyu Guo, Yoshua Bengio, and Shengchao Liu. Assembleflow: Rigid flow matching with inertial frames for molecular assembly. In *The Thirteenth International Conference on Learning Representations*, 2025. 7, 9

[13] Katerina Harvati, Carolin Röding, Abel M Bosman, Fotios A Karakostis, Rainer Grün, Chris Stringer, Panagiotis Karkanas, Nicholas C Thompson, Vassilis Koutoulidis, Lia A Moulopoulos, et al. Apidima cave fossils provide earliest evidence of homo sapiens in eurasia. *Nature*, 571(7766): 500–504, 2019. 2

[14] Andrew D Holland, Jarod M Hutson, Aritza Villaluenga, Tom Sparrow, Andrew Murgatroyd, Alejandro García-Moreno, Elaine Turner, Adrian Evans, Sabine Gaudzinski-Windheuser, and Andrew S Wilson. Digital refit analysis of anthropogenically fragmented equine bone from the schöningen 13 ii-4 deposits, germany. In *Visual Heritage: Digital Approaches in Heritage Science*, pages 305–321. Springer, 2022. 8

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 4

[16] Guillaume Huguet, James Vuckovic, Kilian Fatras, Eric Thibodeau-Laufer, Pablo Lemos, Riashat Islam, Cheng-Hao Liu, Jarrid Rector-Brooks, Tara Akhound-Sadegh, Michael Bronstein, et al. Sequence-augmented se (3)-flow matching for conditional protein backbone generation. *arXiv preprint arXiv:2405.20313*, 2024. 3, 9

[17] Jinhyeok Kim, Inha Lee, and Kyungdon Joo. Fracture assembly with segmentation and iterative registration. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6945–6949. IEEE, 2024. 1

[18] Nikolas Lamb, Cameron Palmer, Benjamin Molloy, Sean Banerjee, and Natasha Kholgade Banerjee. Fantastic breaks: A dataset of paired 3d scans of real-world broken objects and their complete counterparts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4681–4691, 2023. 2, 5, 6, 1

[19] John P Laughlin. *149 Refits: Assessing site integrity and hearth-centered activities at Barger Gulch Locality B*. University of Wyoming, 2005. 1

[20] John P Laughlin and Robert L Kelly. Experimental analysis of the practical limits of lithic refitting. *Journal of Archaeological Science*, 37(2):427–433, 2010. 1, 6

[21] Nahyuk Lee, Juhong Min, Junha Lee, Seungwook Kim, Kanghee Lee, Jaesik Park, and Minsu Cho. 3d geometric shape assembly via efficient point cloud matching. In *Proceedings of the 41st International Conference on Machine Learning*, pages 26856–26873, 2024. 1, 5, 6

[22] Jun Li, Chengjie Niu, and Kai Xu. Learning part generation and assembly for structure-aware shape synthesis. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11362–11369, 2020. 5, 6, 2

[23] Jiahan Li, Chaoran Cheng, Jianzhu Ma, and Ge Liu. Geometric point attention transformer for 3d shape reassembly. *arXiv preprint arXiv:2411.17788*, 2024. 1

[24] Shuangqi Li, Hieu Le, Jingyi Xu, and Mathieu Salzmann. Enhancing compositional text-to-image generation with reliable random seeds. *arXiv preprint arXiv:2411.18810*, 2024. 4

[25] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3, 9

[26] Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky TQ Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code. *arXiv preprint arXiv:2412.06264*, 2024.

[27] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 3, 9

[28] Jiaxin Lu, Gang Hua, and Qixing Huang. Jigsaw++: Imagining complete shape priors for object reassembly. *arXiv preprint arXiv:2410.11816*, 2024. 1

[29] 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*, 2024. 8

[30] Jiaxin Lu, Yifan Sun, and Qixing Huang. Jigsaw: Learning to assemble multiple fractured objects. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3, 5, 6, 7, 8, 2

[31] Pei Luo, Zhuangzhi Wu, Chunhe Xia, Lu Feng, and Teng Ma. Co-segmentation of 3d shapes via multi-view spectral clustering. *The Visual Computer*, 29:587–597, 2013. 8

[32] Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025. 4

[33] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. Peft: State-of-the-art parameter-efficient fine-tuning methods. In *Peft: State-of-the-art parameter-efficient fine-tuning methods*. 2022. 5

[34] Shannon P McPherron, Zeresenay Alemseged, Curtis W Marean, Jonathan G Wynn, Denné Reed, Denis Geraads, René Bobe, and Hamdallah A Béarat. Evidence for stone-tool-assisted consumption of animal tissues before 3.39 million years ago at dikika, ethiopia. *Nature*, 466(7308):857–860, 2010. 1

[35] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 3

[36] Jordy Didier Orellana Figueroa, Jonathan Scott Reeves, Shannon P McPherron, and Claudio Tennie. A proof of concept for machine learning-based virtual knapping using neural networks. *Scientific Reports*, 11(1):19966, 2021. 2, 1

[37] Georgios Papaioannou, E-A Karabassi, and Theoharis Theoharis. Virtual archaeologist: Assembling the past. *IEEE Computer Graphics and Applications*, 21(2):53–59, 2001. 3

[38] Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky TQ Chen. Multisample flow matching: Straightening flows with minibatch couplings. *arXiv preprint arXiv:2304.14772*, 2023. 3, 9

[39] Zipeng Qi, Lichen Bai, Haoyi Xiong, and Zeke Xie. Not all noises are created equally: Diffusion noise selection and optimization. *arXiv preprint arXiv:2407.14041*, 2024. 4

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervi-

sion. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2, 3

[41] Gianluca Scarpellini, Stefano Fiorini, Francesco Giuliari, Pietro Moreiro, 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*, pages 28098–28108, 2024. 1, 3, 5, 8, 9, 2

[42] 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. 1, 2, 5, 6, 8

[43] 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. 2

[44] Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, et al. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*, 2024. 2, 3

[45] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19412–19424, 2024. 2, 3

[46] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017. 3

[47] Theodore Tsesmelis, Luca Palmieri, Marina Khoroshiltseva, Adeela Islam, Gur Elkin, Ofir I Shahar, Gianluca Scarpellini, Stefano Fiorini, Yaniv Ohayon, Nadav Alali, et al. Reassembling the past: The repair dataset and benchmark for real world 2d and 3d puzzle solving. *Advances in Neural Information Processing Systems*, 37:30076–30105, 2025. 1, 2, 5, 8

[48] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 7

[49] Ariana M Villegas-Suarez, Cristian Lopez, and Ivan Sipiran. Matchmakernet: Enabling fragment matching for cultural heritage analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1632–1641, 2023. 1

[50] Lixin Wang and Pauline Sebillaud. The emergence of early pottery in east asia: New discoveries and perspectives. *Journal of World prehistory*, 32:73–110, 2019. 1

[51] Zhengqing Wang, Jiacheng Chen, and Yasutaka Furukawa. Puzzlefusion++: Auto-agglomerative 3d fracture assembly by denoise and verify. *arXiv preprint arXiv:2406.00259*, 2024. 1, 2, 3, 4, 5, 6, 7, 8, 9

[52] Edward O Wilson. *The social conquest of earth*. WW Norton & Company, 2012. 1

[53] Rundi Wu, Yixin Zhuang, Kai Xu, Hao Zhang, and Baoquan Chen. Pq-net: A generative part seq2seq network for 3d shapes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 829–838, 2020. 5, 6, 2

[54] 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*, pages 14311–14320, 2023. 1, 3, 5, 8, 2

[55] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4840–4851, 2024. 3, 5, 1

[56] Jiangyong Xu, Mingquan Zhou, Zhongke Wu, Wuyang Shui, and Sajid Ali. Robust surface segmentation and edge feature lines extraction from fractured fragments of relics. *Journal of Computational Design and Engineering*, 2(2):79–87, 2015. 8

[57] Qun-Ce Xu, Hao-Xiang Chen, Jiacheng Hua, Xiaohua Zhan, Yong-Liang Yang, and Tai-Jiang Mu. Fragmentdiff: A diffusion model for fractured object assembly. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024. 1, 5, 2, 3

[58] Jason Yim, Andrew Campbell, Andrew YK Foong, Michael Gastegger, José Jiménez-Luna, Sarah Lewis, Victor Garcia Satorras, Bastiaan S Veeling, Regina Barzilay, Tommi Jaakkola, et al. Fast protein backbone generation with se (3) flow matching. *arXiv preprint arXiv:2310.05297*, 2023. 3, 9, 2

[59] Jason Yim, Brian L Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. Se (3) diffusion model with application to protein backbone generation. *arXiv preprint arXiv:2302.02277*, 2023. 8

[60] 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. 5, 6, 2

[61] Ruiyuan Zhang, Jiaxiang Liu, Zexi Li, Hao Dong, Jie Fu, and Chao Wu. Scalable geometric fracture assembly via co-creation space among assemblers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7269–7277, 2024. 1

[62] Zikai Zhou, Shitong Shao, Lichen Bai, Zhiqiang Xu, Bo Han, and Zeke Xie. Golden noise for diffusion models: A learning framework. *arXiv preprint arXiv:2411.09502*, 2024. 4

Appendix

This document supplements the main paper as follows:

1. Dataset details (Section A).
2. More details about the training recipe and reproducibility (section B).
3. More visualizations and detailed tables (section C).

A. Additional Dataset Details

A.1. Fracture Simulation

(i) *Bone*. For elongated structures like limbs and ribs, we used Blender's skinning and subdivision surface techniques to create realistic cylindrical hollows, replicating bone morphology. We then applied the physics-based fracture method from Breaking Bad [42] to generate 2–20 fragments. The same approach was used for *os coxae* and vertebrae, forming the simulated subset of the bone category. (ii) *Eggshell*. Since scanned eggshells produce watertight solid ellipsoids, we removed 98% of the concentric volume to simulate thin shells. We then applied the same physics-based fracture method to generate realistic breakage patterns. (iii) *Ceramics*. Given that ceramic objects (e.g., bowls, pots, vases) closely resemble those in Breaking Bad's everyday category, we focused on scanning real fragments and did not include a simulated subset. (iv) *Lithics*. As an initial feasibility test, two generalized core morphologies were repeatedly virtually knapped with some randomized variation following methods described for the dataset in [36] to produce core and flake combinations with varying geometries.

A.2. FRACTURA Statistics

Table I presents detailed statistics for each category in FRACTURA. We continue to expand both the dataset's scale and diversity, aiming to establish a comprehensive cyberinfrastructure for the vision-for-science community.

Table I. Dataset Statistics of the FRACTURA Dataset.

Category	Fracture Type	# Assemblies	# Pieces
Bone	Real	17	37
	Synthetic	7056	39943
Eggshell	Real	3	12
	Synthetic	2268	12600
Ceramics	Real	9	51
	Synthetic	N/A	N/A
Lithics	Real	12	192
	Synthetic	403	807
Total	Real	41	292
	Synthetic	9727	53350

B. Additional Implementation Details

B.1. Data Preprocessing

We preprocess the BreakingBad dataset [42] to calculate the segmentation ground truth directly from meshes to reduce the computation overhead during training as described in Sec. 3.1, and there's no need for any hyperparameters. Unlike baseline methods (Global, LSTM, and DGL) provided by the dataset and PF++ [51], which samples $M = 1000$ points from the mesh per fragment, we used the same setting as in Jigsaw [30] to sample $M = 5000$ points per object, making all fragments have the same point density. With this sampling setting, we did not encounter any gradient explosion issues during training, as reported in FragmentDiff [57], which occur when sampling too many points for tiny pieces. Meanwhile, we employ the Poisson disk sampling method to ensure that the points are more uniformly distributed on the surface of the fragment. During training, standard data augmentation techniques are applied, including recentering, scaling, and random rotation.

B.2. Training Recipe

We modified a smaller version of Point Transformer V3 [55] as our backbone for the segmentation pretraining, as shown in Table II, which we found to be sufficient and more memory efficient. Since GARF uses a much larger training dataset, we reduce the training epochs to 150, other than the 400 epochs used in GARF-mini. Both pretrainings reach over 99.5% accuracy on the validation set. Samples of segmentation results are shown in Fig. I.

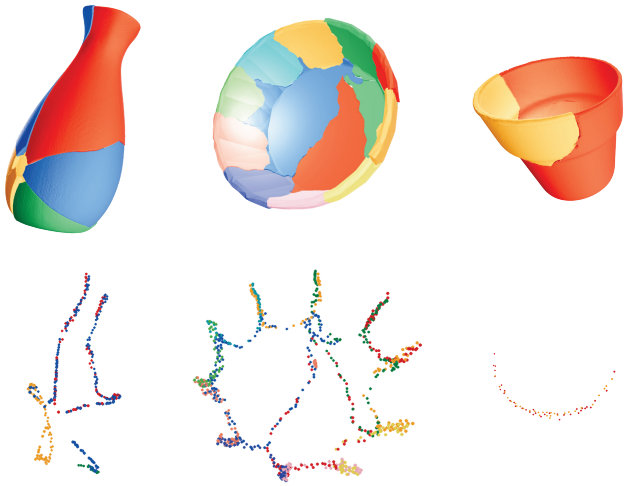

Figure I. Segmentation results on a real-world object (left), Breaking Bad [42] (center) and Fantastic Breaks [18] (right).

For FM training, we provide the hyperparameters in Table II for reproducibility. The settings are identical for both GARF and GARF-mini, as their only difference lies in the

Table II. Training Configurations.

	Config	Value
Backbone	Encoder Depth	[2, 2, 6, 2]
	Encoder # Heads	[2, 4, 8, 16]
	Encoder Patch Size	[1024, 1024, 1024, 1024]
	Encoder Channels	[32, 64, 128, 256]
	Decoder Depth	[2, 2, 2]
	Decoder # Heads	[4, 8, 16]
	Decoder Patch Size	[1024, 1024, 1024]
	Decoder Channels	[256, 128, 64]
Pretraining	Global Batch Size	256
	Epochs	400 / 150
	Learning Rate	1e-4
	Scheduler	CosineAnnealingWarmRestarts
	Scheduler T_0	100 / 50
	# Trainable Params	12.7M
Training	Global Batch Size	128
	Epochs	1500
	Learning Rate	2e-4
	Scheduler	MultiStepLR
	Scheduler Milestones	[900, 1200]
	Scheduler γ	0.5
	# Trainable Params	43.5M

pretraining stage.

B.3. Preliminaries on Riemannian Flow Matching

Instead of simulating discrete noise addition steps, flow matching (FM) learns a probability density path p_t, which progressively transforms a noise distribution $p_{t=0}$ to the data distribution $p_{t=1}$, with a time variable $t \in [0, 1]$. As a simulation-free method aiming to learn continuous normalizing flow (CNF), FM models a probability density path p_t, which progressively transforms a noise distribution $p_{t=0}$ to the data distribution $p_{t=1}$, with a time variable $t \in [0, 1]$. Inspired by *learning assembly by breaking*, the rigid motion of the fragments corresponds to the geodesic on the *Lie group* $\mathrm{SE}(3)$, which is a differentiable Riemannian manifold. Inspired by previous works [4, 11, 58], FM can be extended to $\mathrm{SE}(3)$ manifold to learn the rigid assembly process.

On a manifold $\mathcal{M}$, the flow $\psi_t : \mathcal{M} \to \mathcal{M}$ is defined as the solution of an ordinary differential equation (ODE):

$$\frac{\mathrm{d}}{\mathrm{d}t}\psi_t(\boldsymbol{x}) = \boldsymbol{v}_t(\psi_t(\boldsymbol{x})), \quad \psi_0(\boldsymbol{x}) = \boldsymbol{x}, \tag{8}$$

where $\boldsymbol{v}_t(\boldsymbol{x}) \in \mathcal{T}_{\boldsymbol{x}}\mathcal{M}$ is the time-dependent vector field, and $\mathcal{T}_{\boldsymbol{x}}\mathcal{M}$ is the tangent space of the manifold at $\boldsymbol{x} \in \mathcal{M}$. In the context of $\mathrm{SE}(3)$, the tangent space is the *Lie algebra* $\mathfrak{se}(3)$, which is a six-dimensional vector space, presenting the velocity of the rigid motion of the fragments. Given the *conditional vector filed* $\boldsymbol{u}_t(\boldsymbol{x} \mid \boldsymbol{x}_1) \in \mathcal{T}_{\boldsymbol{x}}\mathcal{M}$, which generates the conditional probability path $p_t(\boldsymbol{x} \mid \boldsymbol{x}_1)$, the Riemannian flow matching objective can be defined as:

$$\mathcal{L}_{\text{CFM}} := \mathbb{E}_{t, p_1(\boldsymbol{x}_1), p_t(\boldsymbol{x}|\boldsymbol{x}_1)} \left[\|\boldsymbol{v}_t(\boldsymbol{x}, t) - \boldsymbol{u}_t(\boldsymbol{x} \mid \boldsymbol{x}_1)\|_G^2 \right], \tag{9}$$

Table III. Results on Vanilla Breaking Bad [42] Dataset.

Methods	RMSE(R) ↓ degree	RMSE(T) ↓ $\times 10^{-2}$	PA ↑ %	CD ↓ $\times 10^{-3}$
Tested on the **Everyday** Subset				
Global [22]	80.70	15.10	24.60	14.60
LSTM [53]	84.20	16.20	22.70	15.80
DGL [60]	79.40	15.00	31.00	14.30
SE(3)-Equiv [54]	79.30	16.90	8.41	28.50
DiffAssemble [41]	73.30	14.80	27.50	-
PHFormer [7]	26.10	9.30	50.70	9.60
Jigsaw [30]	42.30	10.70	57.30	13.30
PF++ [51]	38.10	8.04	70.60	6.03
GARF-mini	**10.41**	**1.91**	**92.77**	**0.45**
Tested on the **Artifact** Subset				
Jigsaw	52.40	22.20	45.60	14.30
PF++	52.10	13.90	49.60	14.50
GARF-mini	**11.91**	**2.74**	**89.42**	**1.05**

where $\|\cdot\|_G^2$ is the norm induced by the Riemannian metric G. Then the learned vector field $\boldsymbol{v}_t$ can be used to generate samples on the manifold at inference, which is $\mathrm{SE}(3)$ poses of the fragments. The rigid motion of fragments corresponds to the geodesic on the *Lie group* $\mathrm{SE}(3)$, a differentiable Riemannian manifold.

C. Additional Results and Analyses

C.1. Quantitative Results on Vanilla Breaking Bad

Given that all our previous experiments were conducted on the volume-constrained version of the Breaking Bad dataset [42], we here provide additional quantitative results on the non-volume-constrained version to align with the settings of previous methods. The results, shown in Table III, demonstrate that our GARF-mini model still significantly outperforms the previous state-of-the-art method, PF++ [51], by a large margin. This performance is consistent across both the everyday and artifact subsets, showcasing the model's robust generalization ability.

We also present the results of FragmentDiff [57] on their custom Breaking Bad dataset in Table IV. FragmentDiff claims to remove tiny pieces, but it is unclear whether this applies only to their training setting or also to evaluation. Unfortunately, since they did not open source their code or provide their preprocessed data, we are unable to directly compare all other methods with FragmentDiff. Additionally, they did not adhere to the common settings used by other methods, which limit the number of pieces from 2 to 20, making direct comparisons on their provided metrics impossible. However, its significant performance drop from the Everyday subset to the Artifact subset suggests that GARF surpasses FragmentDiff in generalization capability.

Table IV. FragmentDiff [57] Results on Their Custom Breaking Bad Dataset.

Methods	Subset	RMSE(R) ↓ degree	RMSE(T) ↓ $\times 10^{-2}$	PA ↑ %
FragmentDiff [57]	Everyday	13.68	7.41	90.20
	Artifact	18.18	8.12	82.30

Table V. Quantitative Per-category Results on the FRACTURA (Synthetic Fracture).

Category	Method	RMSE(R) ↓ degree	RMSE(T) ↓ $\times 10^{-2}$	PA ↑ %	CD ↓ $\times 10^{-3}$
Bone	Jigsaw	66.44	20.54	27.24	91.70
	PF++	66.28	20.50	29.81	47.78
	GARF	17.70	3.80	85.18	5.11
	GARF$_{\text{LoRA}}$	**8.79**	**1.10**	**98.19**	**0.34**
Eggshell	Jigsaw	44.44	12.88	49.03	10.49
	PF++	54.81	13.81	61.36	1.50
	GARF	22.48	6.16	83.41	0.67
	GARF$_{\text{LoRA}}$	**7.10**	**1.95**	**95.68**	**0.26**

C.2. Quantitative Results of Finetuning on the FRACTURA Synthetic Dataset

After finetuning GARF on the FRACTURA synthetic dataset, we report the per-category performance on the bone and eggshell categories, as shown in Table V. The results demonstrate that finetuning the FM model in GARF significantly improves performance on these two unseen categories, showing the effectiveness of our finetuning techniques and the generalizability of our pretraining strategy.

C.3. Additional Qualitative Comparison on the FRACTURA and Breaking Bad Dataset

Figures II, III and IV demonstrate more qualitative comparison on the FRACTURA and Breaking Bad Dataset, where our GARF shows superior performance than the other previous SOTA methods.

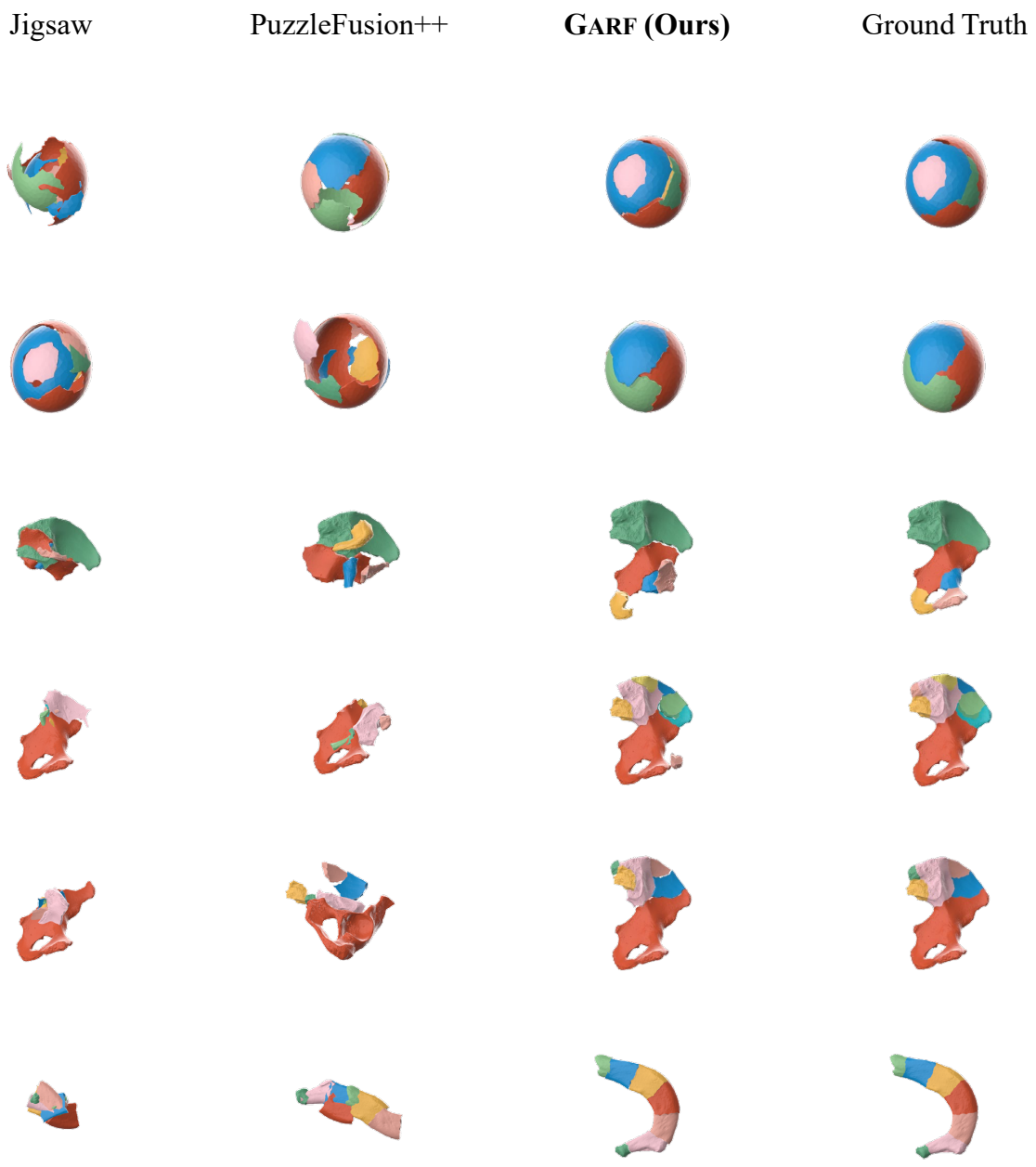

Figure II. Qualitative Results on the FRACTURA Synthetic Dataset.

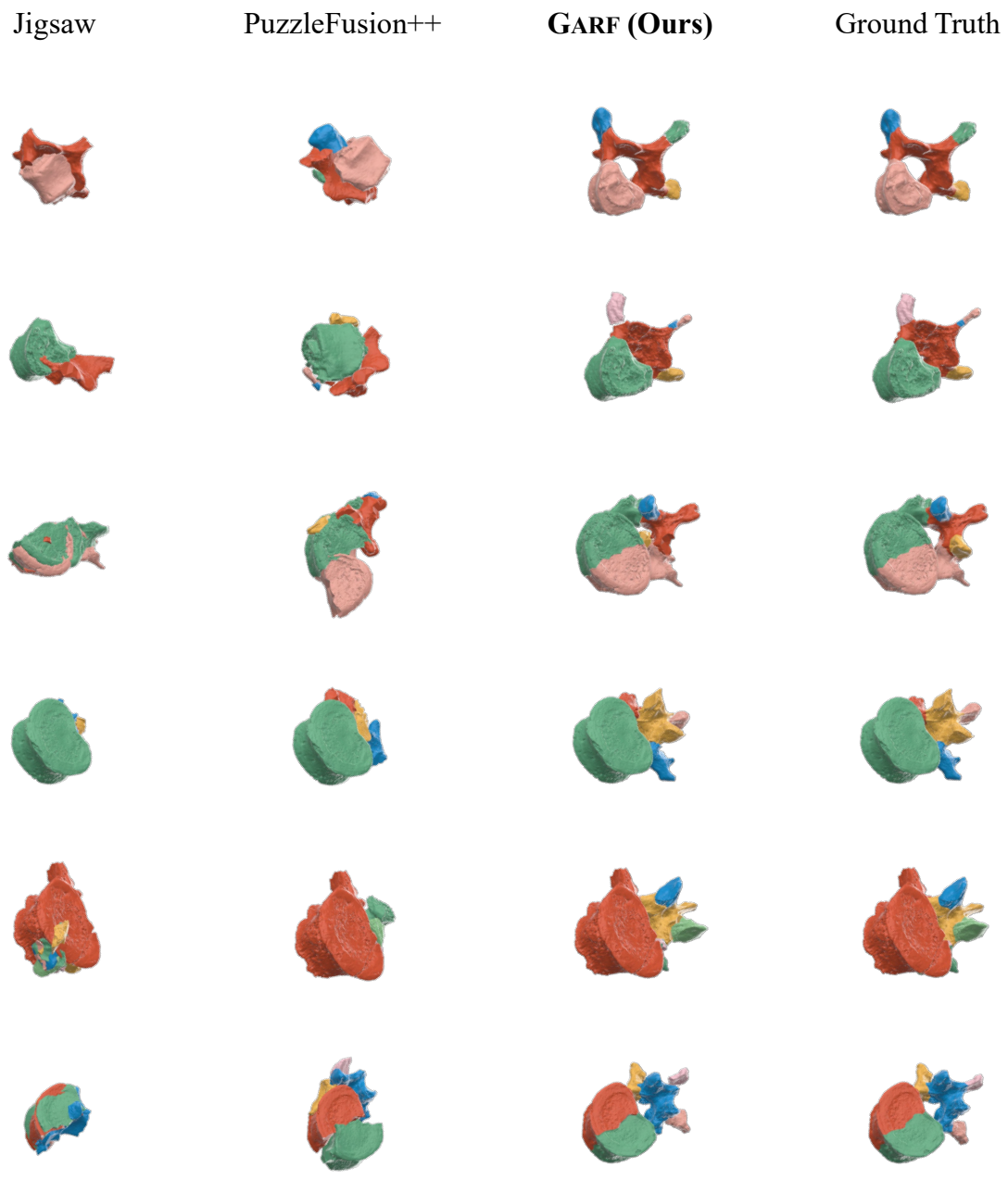

Figure III. Qualitative Results on the FRACTURA Synthetic Dataset.

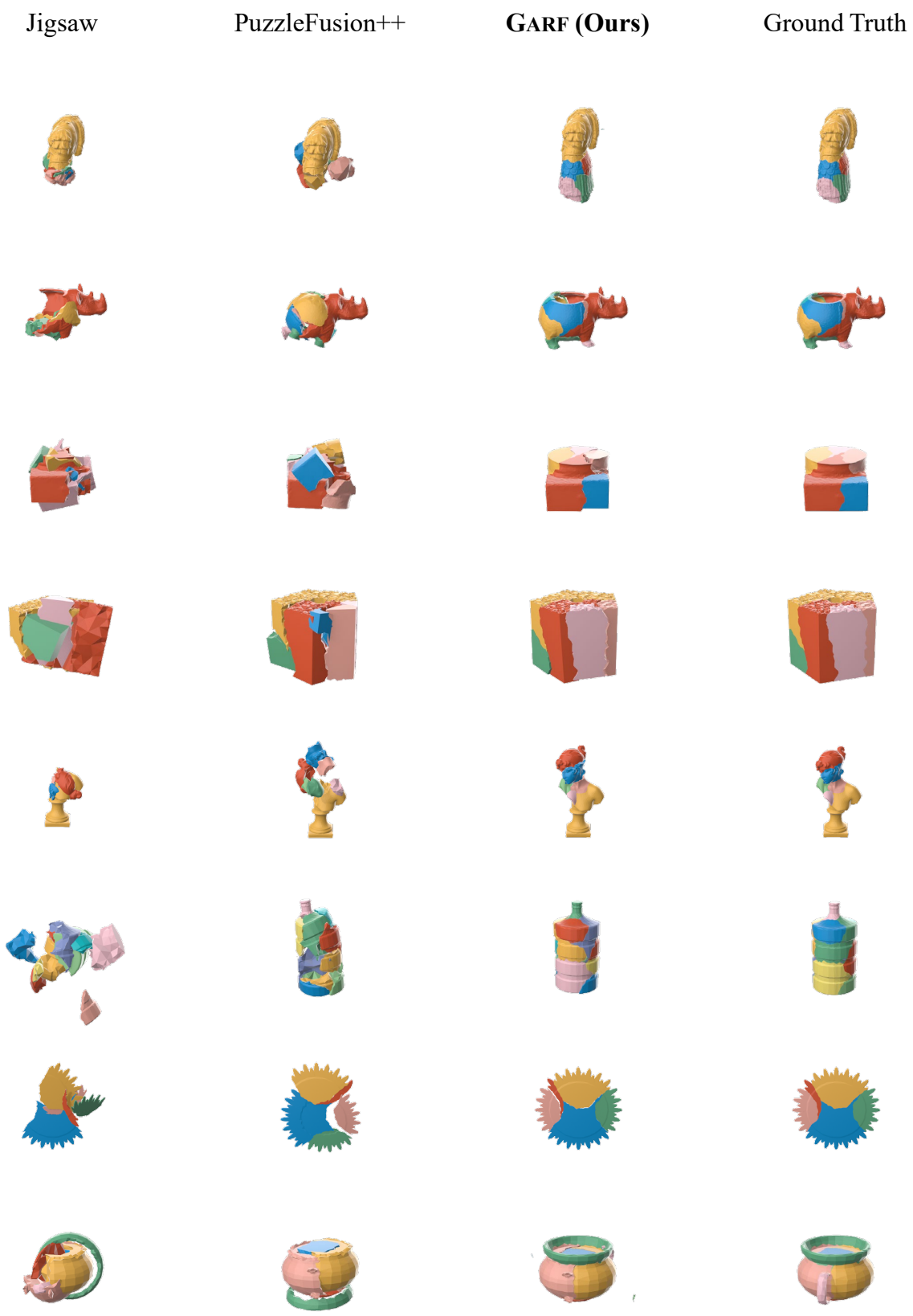

Figure IV. Qualitative Results on the Breaking Bad Dataset Artifact Subset.