CRAG: Can 3D Generative Models Help 3D Assembly?

Zeyu Jiang * **Sihang Li** * **Siqi Tan** † **Chenyang Xu** † **Juexiao Zhang** **Julia Galway-Witham** **Xue Wang**
Scott A. Williams **Radu Iovita** **Chen Feng** ✉ **Jing Zhang** ✉

New York University

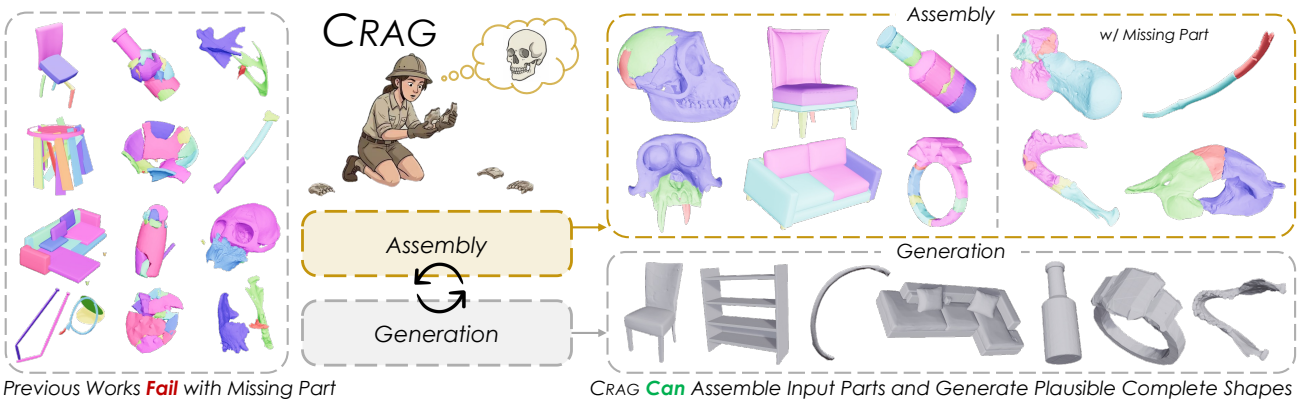

Figure 1: We propose CRAG, a unified framework that couples 3D assembly and generation. CRAG jointly denoises fragment poses and whole-shape latents to assemble the input parts while synthesizing a plausible complete shape, remaining robust to missing parts.

Abstract

Most existing 3D assembly methods treat the problem as pure pose estimation, rearranging observed parts via rigid transformations. In contrast, human assembly naturally couples structural reasoning with holistic shape inference. Inspired by this intuition, we reformulate 3D assembly as a joint problem of assembly and generation. We show that these two processes are mutually reinforcing: assembly provides part-level structural priors for generation, while generation injects holistic shape context that resolves ambiguities in assembly. Unlike prior methods that cannot synthesize missing geometry, we propose CRAG, which simultaneously generates plausible complete shapes and predicts poses for input parts. Extensive experiments demonstrate state-of-the-art performance across in-the-wild objects with diverse geometries, varying part counts, and missing pieces. Our code and models will be released.

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

1. Introduction

3D assembly aims to reconstruct a complete 3D object from a set of parts or fractured fragments, with broad impact in scientific collections (Falcucci et al., 2025; Abrams et al., 2025), medicine (Zhao et al., 2023; Ge et al., 2022), and robot manipulation (Liu et al., 2024). Given partial and noisy observations, an algorithm must infer how pieces align while maintaining global object coherence. In real-world settings, the problem is further complicated by missing, eroded, or partially scanned pieces (Lu et al., 2025b).

Most learning-based approaches formulate 3D assembly as pose estimation, predicting a rigid transformation for each input piece. For example, GARF (Li et al., 2025a) and PuzzleFusion++ (Wang et al., 2025b) use generative models to estimate 6-DoF alignments for rigid fragments. Recently, Assembler (Zhao et al., 2025a) and RPF (Sun et al., 2025) reparameterize 3D assembly in Euclidean space by predicting each part's point set in the assembled state, and then recovering per-part rigid transformations via least-squares fitting or Procrustes (Gower, 1975) solved with SVD. Assembler (Zhao et al., 2025a) additionally conditions on an image by injecting pretrained visual features to provide global shape cues. However, these approaches are designed to transform the geometry of the observed parts, so they primarily reposition input points and do not synthesize new geometry to fill missing regions.

In contrast, human experts do not treat assembly as local alignment alone. They iteratively hypothesize the unseen whole while placing fragments, using a progressively refined global shape hypothesis to resolve ambiguity (Wagner, 2013). This is especially critical when pieces are missing, where conservators often infer absent regions from the available fragments and perform gap-filling to restore a plausible, complete form. Motivated by this perspective, we ask an open question: *could 3D assembly and generation be unified to benefit each other?*

Achieving such a unification raises two core challenges. First, one must construct a shared latent space that can embed an unordered set of variable-size, partial, and noisy fragments into a representation that is simultaneously usable for assembly and compatible with a generative shape prior, so that gradients and uncertainty can flow across the two branches rather than being trapped in mismatched latent spaces. Second, even with a shared representation, one must enable *effective bidirectional information exchange*: fragment evidence should steer what the whole should be, while the imagined whole should in turn disambiguate how fragments should align, without creating unstable feedback loops during joint training and inference.

We address these challenges with CRAG, which Couples ReAssembly and Generation in a *joint flow-matching* framework. Concretely, CRAG performs joint denoising by simultaneously predicting an SE(3) flow for fragment poses and a latent-space flow for whole-shape generation. To establish a common "language" between the two tasks, CRAG reuses the VAE from TripoSG (Li et al., 2025b) as a shared embedding space, mapping variable-size fragment sets into features that live in the same latent space as whole-shape generation and are therefore compatible with a strong generative prior. Building on this shared space, CRAG adopts a Mixture-of-Transformers architecture with two parallel branches and introduces a Joint Adapter at each layer. The key design of the Joint Adapter is the bi-directional attention mechanism to enable mutual refinement: fragment features inform what the whole should be, while the imagined whole guides how fragments should align. This design mirrors human experts, who iteratively reconcile fragment observations with hypotheses of the unseen whole. For stable learning, we employ a two-stage training strategy, learning assembly first and then jointly finetuning both tasks.

Our main contributions are as follows:

- **A New Capability for 3D Assembly.** CRAG jointly assembles fragments and synthesizes plausible complete shapes, remaining robust to missing parts while improving alignment under ambiguity.
- **A New Formulation and Framework.** We recast 3D assembly as a coupled reassembly-and-generation objective, and propose a joint flow-matching framework that denoises fragment poses in SE(3) and whole-shape latents in a single inference loop.
- **SOTA Results and a New Dataset.** CRAG achieves SOTA performance on part assembly (PartNeXt (Wang et al., 2025a)) and fracture reassembly (Breaking Bad (Sellán et al., 2022)). To facilitate future research, we curate and release a new bone fragment dataset from MorphoSource (Boyer et al., 2016).

2. Related Work

3D Assembly. 3D assembly is a long-standing problem that mirrors human spatial intelligence: it requires reconciling local cues with a coherent global hypothesis under ambiguity. Most methods formulate it as pose estimation, predicting an SE(3) transform for each piece (Lee et al., 2024). Early methods relied on hand-crafted descriptors that generalize poorly. Learning-based methods such as Jigsaw (Lu et al., 2023) and Combinative Matching (Lee et al., 2025) learn correspondences, improving robustness. Recent approaches including GARF (Li et al., 2025a), PuzzleFusion++ (Wang et al., 2025b), and DiffAssemble (Scarpellini et al., 2024) employ generative models that iteratively denoise poses on SE(3). Another line of work leverages complete-shape priors to further constrain this challenging problem. Classical priors rely on explicit templates, e.g., symmetry (Koutsoudis et al., 2010), while recent methods such as RPF (Sun et al., 2025) and Assembler (Zhao et al., 2025a) predict assembled point sets in Euclidean space and recover rigid transforms via SVD. Assembler further uses a reference image to resolve ambiguity, a capability that CRAG also supports through an optional image condition. Jigsaw++ (Lu et al., 2025a) takes a partially assembled point cloud and applies a learned complete-shape prior, but this separated design limits mutual refinement. In contrast, CRAG takes a first step toward jointly coupling assembly and generation within a unified optimization loop, enabling bidirectional benefits.

3D Generation for Assembly. Recent advances in 3D generation provide powerful shape priors that can benefit 3D assembly. These models are typically built upon two-stage designs that learn a compact VAE latent space and then model it with diffusion or flow-based transformers (Wu et al., 2024a), often using efficient VecSet latents (Zhang et al., 2023). Representative open models such as Dora (Chen et al., 2025), TripoSG (Li et al., 2025b), and Hunyuan3D 2.0 (Zhao et al., 2025b) follow this paradigm and achieve strong image-to-3D performance by conditioning the latent denoiser on image features. In parallel, TRELLIS (Xiang et al., 2025) introduces a unified structured latent that can be decoded into multiple 3D formats, such as 3D Gaussians and meshes. Beyond image or language prompts, Hunyuan3D-Omni (Hunyuan3D et al., 2025) augments conditioning with point clouds, voxels, 3D bounding boxes, and

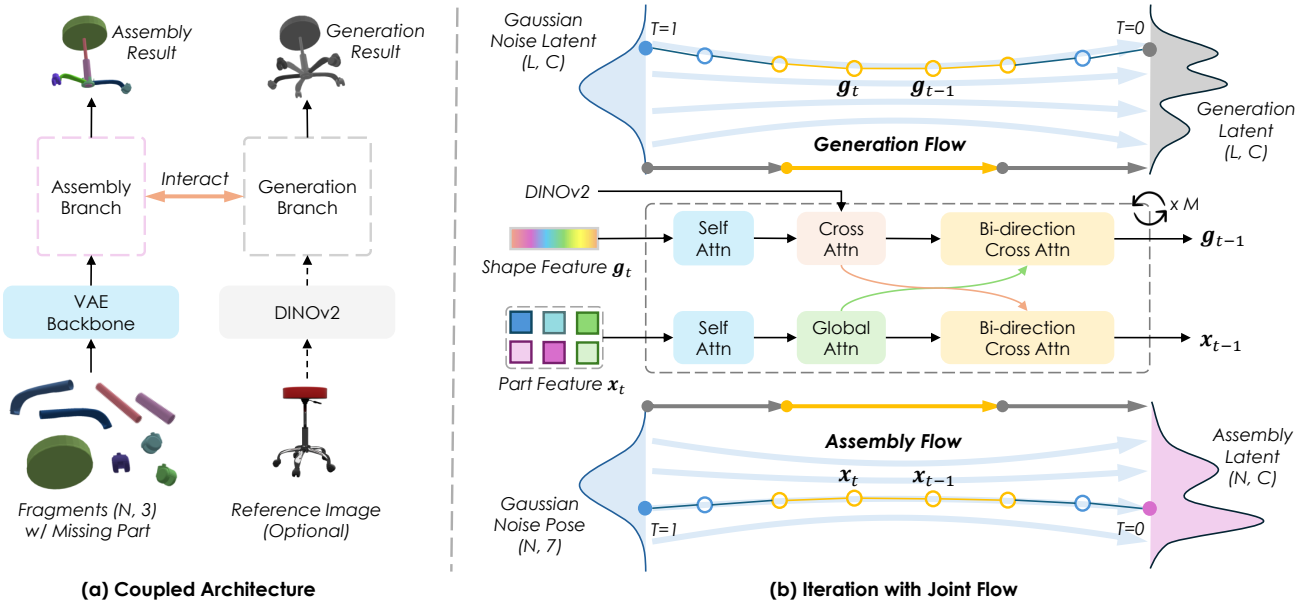

Figure 2. Overview of our approach CRAG . We propose a unified framework for 3D assembly and whole-shape generation. Our model consists of two interacting branches: an *Assembly Branch* that predicts the pose for each part via SE(3) flow matching, and a *Generation Branch* that synthesizes the complete shape via flow matching. A *Joint Adapter* bridges these branches, enabling bidirectional information flow. We employ a two-stage training strategy: learning assembly first, and then jointly finetuning both tasks.

skeletal pose priors for fine-grained controllability. CRAG reuses TripoSG's Transformer VAE as a shared VecSet latent space and pre-trained weights as the generation branch, while our joint flow matching allows assembled parts to provide additional structural priors when image cues are ambiguous.

3. Method

Given a set of fragments $\mathcal{P} = \{p_i\}_{i=1}^{N}$ and an optional reference image I, our goal is to simultaneously recover the pose $\boldsymbol{T}_i \in \mathrm{SE}(3)$ for each fragment and reconstruct the complete 3D shape $\mathcal{S}$. To achieve this, we propose a unified framework CRAG, as illustrated in Figure 2. In the following subsections, we first review the pre-trained generative backbone, TripoSG (Li et al., 2025b) (§3.1), then describe how we leverage its VAE for fragment embedding (§3.2), and finally introduce our joint flow architecture for pose estimation and shape generation (§3.3).

3.1. Preliminary: TripoSG

We first briefly review TripoSG (Li et al., 2025b), the foundational image-to-3D model for our method. TripoSG is a flow-matching-based generative model that operates on a latent set $z \in \mathbb{R}^{n\times d}$, consisting of n latent vectors of dimension d. The latent representation is encoded by a transformer-based VAE following the strategy introduced in 3DShape2VecSet (Zhang et al., 2023). Given a 3D shape $\mathcal{S}$, TripoSG first samples M surface points $P \in \mathbb{R}^{M\times 3}$ from $\mathcal{S}$, and then draws a subsampled point set $P' \subset P$. Both point sets are embedded using a Fourier feature embedding $\gamma(\cdot)$. The initial latent set z is computed by stacking multiple multi-head self-attention (SA) layers on top of a multi-head cross-attention module:

$$z_0 = \mathrm{CrossAttn}\big(\gamma(P'), \gamma(P)\big), \tag{1}$$

$$z = \mathrm{Linear}\big(\mathrm{SelfAttn}_i(z_0)\big), \quad i \in [1, L^E_{\mathrm{SA}}], \tag{2}$$

where L^E_{SA} denotes the number of SA layers in the encoder. The output z serves as the noise-adding target for the denoiser model. Afterwards, z can be decoded by another set of self-attention layers, followed by a cross-attention to decode the signed distance function (SDF) values at the query points x.

$$\tilde{z} = \mathrm{Linear}\big(\mathrm{SelfAttn}_j(z)\big), \quad j \in [1, L^D_{\mathrm{SA}}], \tag{3}$$

$$\tilde{s} = \mathrm{CrossAttn}\big(\gamma(x), \tilde{z}\big), \tag{4}$$

where L^D_{SA} denotes the number of SA layers in the decoder.

3.2. Shared VAE for Fragment Embedding

Why Using Shared VAE. Most recent fragment assembly methods learn fragment-level representations by solving task-specific binary classification problems, such as predicting fracture surfaces (Li et al., 2025a) or detecting fragment overlaps (Sun et al., 2025). However, such encoders are inherently limited by the fragmentation patterns present in their training data and often fail to generalize to unseen fragmentations. In contrast, we propose to reuse the VAE

from TripoSG (Li et al., 2025b) as a *shared fragment embedding module*. Since this VAE is trained on large-scale 3D shape datasets, it provides a more robust and generalizable geometric prior. More importantly, sharing the same VAE between part assembly and whole-shape generation establishes a unified latent space, which facilitates effective information exchange between the two branches.

Fragment Encoding. To support variable numbers of fragments and uneven fragment sizes, we first adapt the VAE encoder with a customized attention processor that allows variable-length input. Given a set of fragments $\mathcal{P} = \{p_i\}_{i=1}^{N}$, we sample M_i points proportional to each fragment's area, then downsample by a factor of 4 to form the VAE query set P_i'. Each fragment is then independently encoded using the shared VAE encoder, producing a fragment-specific latent feature. Instead of directly using encoded latent z as the fragment representation, we use the decoded latent $\tilde{z}$ as the fragment feature. We found that this choice provides better alignment between the assembly and generation branches and yields better performance.

3.3. Joint Flow Matching

The proposed joint flow matching adopts a Mixture-of-Transformers architecture that coordinates the assembly and generation processes. As shown in Figure 2, the model consists of two parallel Transformer branches: an *Assembly Branch* that predicts the per-fragment SE(3) flow, and a *Generation Branch* that predicts the flow for the shape latents. A *Joint Adapter* serves as a bridge, enabling bidirectional information flow between the two branches at each layer.

Assembly Branch. Following GARF (Li et al., 2025a), we model the assembly process as a continuous flow on the manifold $\mathcal{M} = \mathrm{SO}(3) \times \mathbb{R}^3$. Given an initial state $(\boldsymbol{r}_0, \boldsymbol{a}_0)$ and a target state $(\boldsymbol{r}_1, \boldsymbol{a}_1)$, the flow trajectory is defined as:

$$\begin{aligned} \boldsymbol{r}_t &= \exp_{\boldsymbol{r}_0}\left(t \log_{\boldsymbol{r}_0}(\boldsymbol{r}_1)\right), \\ \boldsymbol{a}_t &= (1-t)\boldsymbol{a}_0 + t\boldsymbol{a}_1, \end{aligned} \tag{5}$$

where $t \in [0, 1]$. Here, $\boldsymbol{r}$ represents orientation (handled via geodesics on SO(3)) and $\boldsymbol{a}$ represents position (linear interpolation). We train the flow matching network by minimizing the difference between the predicted velocity and the ground-truth vector field:

$$\begin{aligned} \mathcal{L} = \mathbb{E}_{t,(\boldsymbol{r}_1,\boldsymbol{a}_1)}\Bigg[\sum_{i=1}^{N}\Big(&\left\|f_r^i(\boldsymbol{r}_t, \boldsymbol{a}_t, t) - u_t\right\|^2 \\ &+ \left\|f_a^i(\boldsymbol{r}_t, \boldsymbol{a}_t, t) - v_t\right\|^2\Big)\Bigg], \end{aligned} \tag{6}$$

where the flow targets are defined as $u_t = \frac{\log_{\boldsymbol{r}_t}(\boldsymbol{r}_1)}{1-t}$ and $v_t = \frac{\boldsymbol{a}_1 - \boldsymbol{a}_t}{1-t}$. We omit the fragment index i for brevity.

Generation Branch. While the assembly branch operates on fragment-level poses, the generation branch works in the latent space to reconstruct the complete shape. Following TripoSG (Li et al., 2025b), we model the generation process as a continuous-time flow between the clean latent $z_0 \in \mathbb{R}^{n \times d}$ (encoded by the shared VAE) and Gaussian noise $z_1 \sim \mathcal{N}(0, I)$. At each time step $t \in [0, 1]$, the generation transformer predicts the velocity field:

$$v_z = f_{\text{gen}}(z_t, t, c_I), \tag{7}$$

where c_I is an optional image condition extracted from the reference image I using a frozen DINOv2 (Oquab et al., 2023) encoder. When no reference image is provided, c_I is set to zero. The training objective minimizes MSE between predicted and ground-truth velocities $\mathcal{L}_{\text{gen}} = \mathbb{E}[\|v_z - (z_0 - z_1)\|^2]$. During inference, we iteratively denoise from z_1 to z_0 using the Euler method, and decode the final shape $\mathcal{S}$ using the shared VAE decoder.

Importantly, the generation transformer is initialized from the pre-trained TripoSG (Li et al., 2025b) model and can be either frozen or fine-tuned, allowing us to leverage powerful geometric priors from large-scale 3D datasets while optionally adapting to fragmented objects.

Joint Adapter. A key challenge in our unified framework is enabling effective information exchange between the assembly and generation branches. While both branches condition on the global image context c_I, they operate on fundamentally different representations: the assembly branch processes variable-length fragment point clouds, whereas the generation branch works with fixed-size latent tokens. To bridge this gap, we introduce a *Joint Adapter* module at each transformer layer.

Let $h_{\text{asm}}^{(\ell)} \in \mathbb{R}^{M \times d}$ and $h_{\text{gen}}^{(\ell)} \in \mathbb{R}^{L \times d}$ denote the hidden states of the assembly and generation branches at layer ℓ, respectively. The Joint Adapter facilitates bidirectional information flow through a symmetric cross-attention mechanism. Formally, the hidden states are updated as:

$$\begin{aligned} h_{\text{gen}}^{(\ell)} &= h_{\text{gen}}^{(\ell)} + \text{CrossAttn}\big(h_{\text{gen}}^{(\ell)}, h_{\text{asm}}^{(\ell)}\big), \\ h_{\text{asm}}^{(\ell)} &= h_{\text{asm}}^{(\ell)} + \text{CrossAttn}\big(h_{\text{asm}}^{(\ell)}, h_{\text{gen}}^{(\ell)}\big), \end{aligned} \tag{8}$$

where $\text{CrossAttn}(x, y)$ computes the multi-head attention with x serving as the query and y as the key/value context. This formulation allows the generation branch to query geometric details from the fragments, while the assembly branch simultaneously incorporates structural priors from the generation latent space. Crucially, to ensure stable training and preserve the pre-trained priors of the generation model, we initialize the output projection layers of the adapter with zero weights. This ensures that the adapter acts as an identity mapping at the beginning of training.

Table 1. Quantitative comparison on PartNeXt (Wang et al., 2025a) and Breaking Bad (Sellán et al., 2022) under two part-status settings: *Complete* (all parts observed) and *Missing* (with missing parts). CRAG consistently achieves the best overall performance across datasets and remains robust in the challenging missing-part setting. The **best** and second best results are highlighted.

	PartNeXt (Wang et al., 2025a)								Breaking Bad (Sellán et al., 2022)							
Part Status	Complete				Missing				Complete				Missing			
Method	RE ↓	TE ↓	PA ↑	CD ↓	RE ↓	TE ↓	PA ↑	CD ↓	RE ↓	TE ↓	PA ↑	CD ↓	RE ↓	TE ↓	PA ↑	CD ↓
GARF (Li et al., 2025a)	52.52	10.68	58.19	3.10	49.40	9.73	60.71	5.78	8.75	1.72	93.56	0.42	14.59	3.01	85.55	3.43
RPF (Sun et al., 2025)	54.99	29.54	46.17	10.46	42.49	22.95	57.20	8.21	30.59	13.67	80.23	1.01	31.04	14.96	77.05	1.20
CRAG w/o img (Ours)	45.45	9.82	61.67	3.31	42.46	8.70	66.74	5.17	**8.00**	1.37	94.64	0.22	**11.43**	1.79	92.03	0.52
Assembler (Zhao et al., 2025a)	85.82	17.14	44.18	27.93	83.80	15.00	49.86	24.71	74.92	13.22	48.38	7.46	74.45	13.43	45.93	7.78
CRAG (Ours)	**45.12**	**9.13**	**65.89**	**2.40**	**42.33**	**7.86**	**71.81**	**4.21**	**8.00**	**1.36**	**94.68**	**0.21**	11.44	**1.77**	**92.07**	**0.50**

4. Experiment

4.1. Implementation Details

CRAG is built upon the pretrained TripoSG (Li et al., 2025b) and adopts its pretrained VAE for part-level feature extraction. The assembly branch follows GARF's (Li et al., 2025a) design, but extends to 21 layers to match the depth of the pretrained generation branch with skip connections. To stabilize training and accelerate convergence, we employ a two-stage training strategy. In the first stage, only the assembly branch is trained for 100k steps for warm-up. In the second stage, the generation branch is activated with joint adapters, and the entire model is jointly trained for an additional 150k steps. We keep the classifier-free guidance strategy in the second stage, but increase the image condition drop rate to 50% from 10% along the training process to enable the generation branch to learn from the part-level assembly information. The full training process takes about 3 days on 32 NVIDIA H200 GPUs, with a global batch size of 256 in the first stage and 128 in the second stage.

4.2. Experiment Setup

Datasets. We define two tasks, *part assembly* and *fracture reassembly*, and train and evaluate our method on each task separately. (i) *Part Assembly*. We use **PartNeXt** (Wang et al., 2025a) as both the training dataset and evaluation benchmark. PartNeXt contains 23,519 textured 3D models, from which we sample 15,563 shapes for training and 3,903 shapes for evaluation, restricting the number of parts per shape to the range of 2 to 20. PartNeXt focuses on *semantic parts* (e.g., chair legs and table tops), which are defined by functional or semantic boundaries rather than geometric fracture surfaces. (ii) *Fracture Reassembly*. For this task, we use the everyday subset of the **Breaking Bad** (Sellán et al., 2022) dataset, combined with a curated collection from **MorphoSource** (Boyer et al., 2016). MorphoSource is an open-access repository containing thousands of 3D models of primate (including human) bones and bones from other animals. We virtually fracture these models using Breaking Good (Sellán et al., 2023), FractureRB (Hahn & Wojtan, 2016), and FractureBEM (Hahn & Wojtan, 2015) to substantially increase sample size and taxonomic coverage. The resulting dataset contains 3,347 samples across 10 categories. Complementarily, we use the **FRACTURA** dataset (Li et al., 2025a) to validate CRAG on real-world fractures.

Evaluation Metrics. Following (Li et al., 2025a; Sun et al., 2025; Wang et al., 2025b), we evaluate the assembly quality by following metrics: (i) **Rotation Error (RE)** is the root mean square error of Euler angles for each part. (ii) **Translation Error (TE)** measures the root mean square error of translation vectors for each part. (iii) **Part Accuracy (PA)** is the fraction of parts whose the chamfer distance to the ground truth is less than 10^{-2}. (iv) **Chamfer Distance (CD)** computes the chamfer distance between the assembled shape and the ground truth shape. All metrics are averaged over all parts and all shapes.

Baseline Methods. We compare our approach against state-of-the-art methods for 3D part assembly. **GARF** (Li et al., 2025a) and **RPF** (Sun et al., 2025) are point cloud-based methods that take only part geometry as input. GARF directly regresses the SE(3) pose for each part, while RPF first predicts the target location of each part's point cloud and subsequently recovers the rigid transformation through SVD optimization. **Assembler** (Zhao et al., 2025a) is the most related work to ours, which augments the input with multi-view images to provide additional visual context. Its overall pipeline follows a similar predict-then-optimize paradigm as RPF. We denote our method as **CRAG** and report results on both datasets described above. Additionally, we extend the evaluation protocol to include scenarios with *missing parts* during inference.

4.3. Evaluation

We conduct extensive experiments to address three central questions of this work:

Q1: Does generation improve assembly by providing a holistic prior?

Q2: Can we still assemble the observed parts and synthesize a complete object from incomplete fragment sets?

Q3: Does part-level evidence help disambiguate image-conditioned 3D generation?

Figure 3. Qualitative results across PartNeXt (Wang et al., 2025a), Breaking Bad (Sellán et al., 2022), and MorphoSource (Boyer et al., 2016). We first compare methods without reference images by contrasting GARF (Li et al., 2025a), RPF (Sun et al., 2025), and CRAG w/o image, where CRAG produces more coherent assemblies and more complete shapes from the same observed parts. We then compare image-conditioned methods by showing Assembler and full CRAG given the reference image, where CRAG better aligns parts and yields shapes that more closely match the ground truth.

Holistic Shape Priors Boost Assembly Performance. Table 1 reports quantitative results on PartNeXt (Wang et al., 2025a) and Breaking Bad (Sellán et al., 2022) under the complete-part setting. We evaluate two variants of our approach: **CRAG**, which uses a reference image, and **CRAG w/o img**, which uses only part point clouds. To answer **Q1** fairly, we first compare methods without reference images, where CRAG w/o img consistently outperforms GARF (Li et al., 2025a) and RPF (Sun et al., 2025) across datasets, indicating that coupling assembly with generation provides a holistic prior that regularizes pose recovery from geometry alone. When reference images are available, CRAG further surpasses the image-conditioned Assembler (Zhao et al., 2025a) by a large margin. On PartNeXt, CRAG increases PA from 44.18 to 65.89 and reduces CD from 27.93 to 2.40, corresponding to a 91.4% relative reduction.

These quantitative trends are also evident in Figure 3. GARF (Li et al., 2025a) and RPF (Sun et al., 2025) often produce unstable pose estimates under ambiguity, leading to visibly inconsistent part placements such as tilted or floating structures on PartNeXt (Wang et al., 2025a) and severe misalignments on Breaking Bad (Sellán et al., 2022) and MorphoSource (Boyer et al., 2016). In contrast, CRAG w/o img yields coherent assemblies across all three datasets, preserving global structure and maintaining consistent part-to-part contacts, which suggests that the coupled generation prior provides a strong holistic constraint beyond local geometric cues. When conditioned on the reference image, Assembler (Zhao et al., 2025a) can reduce some ambiguities but still exhibits implausible global structure or residual part misplacement, whereas full CRAG produces assemblies that more closely match the ground truth and simultaneously synthesizes more complete and globally consistent shapes. Overall, these results answer **Q1** affirmatively:

> **A1:** *The generation prior provides holistic structural guidance that improves assembly.*

Robust Assembly and Shape Synthesis with Missing Parts. In the missing-part setting of Table 1, CRAG remains robust even without reference images. CRAG w/o img consistently outperforms GARF (Li et al., 2025a) and RPF (Sun et al., 2025) on both datasets, with higher part accuracy and lower shape error, showing stronger assembly performances from incomplete observations. On PartNeXt (Wang et al., 2025a), it increases PA to 66.74 while reducing CD to 5.17, improving over GARF and RPF. On Breaking Bad (Sellán et al., 2022), the gap is larger, reaching 92.03 PA with 0.52

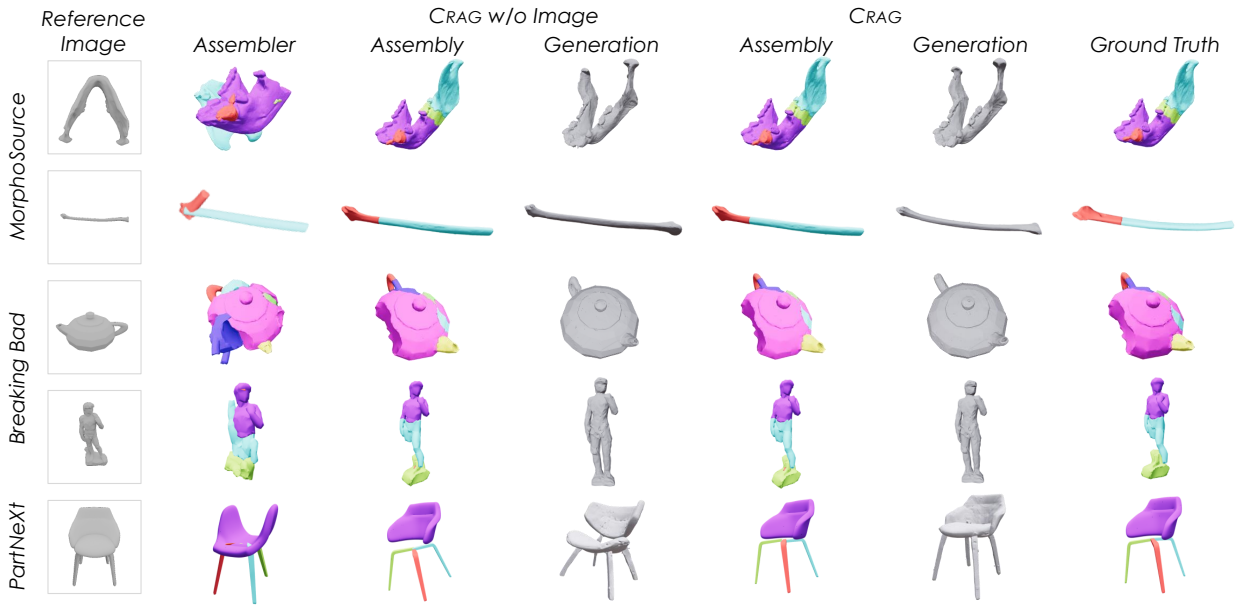

Figure 4. Qualitative results on PartNeXt (Wang et al., 2025a), Breaking Bad (Sellán et al., 2022), and MorphoSource (Boyer et al., 2016) with missing parts. We compare Assembler, CRAG without reference images, and CRAG given a reference image. CRAG simultaneously assembles the observed parts and synthesizes a plausible, complete shape, and reference images further improve fidelity when available.

CD, substantially surpassing both baselines under severe incompleteness. When reference images are available, CRAG further improves and clearly exceeds the image-conditioned Assembler (Zhao et al., 2025a), which degrades sharply with missing parts. Figure 4 visualizes CRAG's joint assembly-and-generation behavior: it not only places the available parts coherently but also synthesizes a globally consistent complete shape by hallucinating the missing geometry, with reference images further improving the fidelity of the generated completion when available. These answer **Q2**:

> **A2:** CRAG *can robustly assemble the observed parts and simultaneously synthesize a plausible complete object, even when parts are missing.*

Part-Level Evidence Helps Disambiguate Image-Conditioned Generation. Our joint flow architecture enables a novel capability: resolving the inherent ambiguity in image-based 3D generation. Existing methods struggle to infer occluded regions from a single view, often producing geometrically inconsistent reconstructions. Qualitative results in Figure 5 demonstrate that our approach produces coherent completions even for heavily occluded objects. Part-level evidence can, to some extent, help disambiguate image-conditioned 3D generation. When the reference view is ambiguous or incomplete, for example, when the camera viewpoint does not reveal the full object, TripoSG (Li et al., 2025b) may produce inconsistent shapes, while CRAG uses the assembled parts as additional evidence to resolve some ambiguities and generate shapes closer to the ground truth. These observations answer **Q3** in a conservative way:

Figure 5. Qualitative results under ambiguous reference images on PartNeXt (Wang et al., 2025a). We compare image-only generation with TripoSG (Li et al., 2025b) against CRAG by visualizing CRAG's assembled parts and generated shapes alongside the ground truth. When the reference view is incomplete and does not reveal the full object, part-level evidence helps, to some extent, resolve ambiguity and yields a better shape.

> **A3:** *To some extent, part-level evidence can reduce ambiguity in image-conditioned 3D generation by providing additional geometric constraints when the reference view is incomplete.*

Ablation Study. Table 2 ablates key design choices in CRAG on PartNeXt (Wang et al., 2025a). **(i) Fragment Encoding.** ① uses a PTv3 (Wu et al., 2024b) encoder with GARF-style task-specific pretraining (e.g., fracture-surface segmentation), while ② replaces it with our adapted TripoSG VAE embedding (decoded $\tilde{z}$). The consistent gains in ② suggest that large-scale 3D generative pretraining provides a stronger geometric prior for assembly. **(ii) Genera-**

Table 2. Ablation study (PartNeXt (Wang et al., 2025a), Complete) with key design choices. "A" and "G" refer to the assembly branch and the generation branch, respectively.

Setups	Enc.	Ref. Img	Branch	RE ↓	TE ↓	PA ↑	CD ↓
①	PTv3	×	A	52.52	10.68	58.19	3.10
②	VAE	×	A	47.16	10.12	60.01	3.61
③	VAE	✓	A	47.13	9.60	62.29	2.69
④	VAE	×	A + G	45.45	9.82	61.67	3.31
⑤ (CRAG)	VAE	✓	A + G	**45.12**	**9.13**	**65.89**	**2.40**

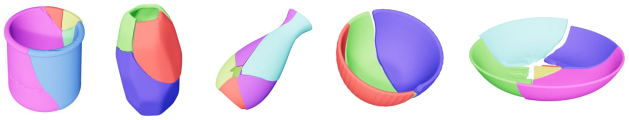

Figure 6. Qualitative results of CRAG on FRACTURA (Li et al., 2025a), demonstrating robustness on real-world fractures. All parts are real scanned fragments; colors are rendered for visualization.

tion Branch. To quantify how much the generation branch helps 3D assembly beyond simply adding an image condition, we compare assembly-only with images ③ against the full coupled model ⑤. The gain from ③ to ⑤ indicates that coupling assembly with generation provides additional holistic shape context that resolves ambiguities not captured by image features alone. (**iii**) **Robustness without Reference Images.** Comparing ④ against ⑤, together with the strong performance of ④, shows that CRAG remains effective even without a reference image: the generative prior can still regularize and guide pose denoising, while images further boost performance when available.

Real-World Fractures. In Figure 6, we visualize CRAG on the FRACTURA (Li et al., 2025a) to validate performance on real-world fractures.

Representative Failure Cases. In Figure 7, we observe three representative failure modes of CRAG. First, thin shell fragments provide weak pose constraints, so the assembly flow may converge to a plausible but incorrect SE(3) alignment. Second, the TSDF-based VAE encoding underrepresents slender components when their thickness falls below the truncation distance, causing opposite surfaces to merge and producing broken structures. Third, for non-watertight surfaces, signed distance is ambiguous near boundaries, biasing reconstructions toward watertight shapes with hole closing or sheet thickening, and these errors propagate through the assembly generation coupling.

5. Conclusion and Discussion

We present CRAG, a joint flow-matching framework that couples 3D reassembly and whole-shape generation in the wild. CRAG reuses the TripoSG VAE and its large-scale pretrained generative weights to anchor a shared latent space, then jointly denoises fragment poses on SE(3) and whole-shape latents, with a Joint Adapter enabling bidirectional

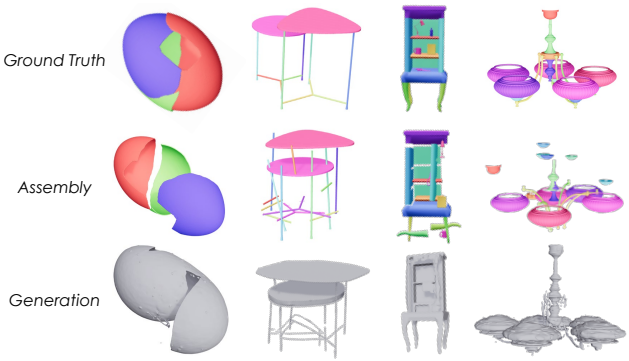

Figure 7. Our representative failure cases.

information exchange between the two branches. Across extensive benchmarks, we find that this coupling injects a holistic shape context that improves assembly under missing-part settings, while part-level evidence helps disambiguate image-conditioned generation.

Broader Impacts. Across archaeology and paleoanthropology, 3D assembly transforms fragmented artifacts, bones, and fossil scans into coherent digital specimens, enabling scalable and reproducible morphometric analysis and cross-site comparison beyond what manual refitting can support. CRAG can further speed hypothesis testing on anatomy and evolutionary variation, particularly for incomplete or eroded specimens. In medicine, analogous multi-fragment reconstruction from CT supports preoperative assessment, surgical planning, and reduction guidance, complementing robot-assisted fracture reduction that seeks higher accuracy and safety. In robotics, everyday repair and maintenance require reasoning about spatial part relations under occlusion and ambiguity, and CRAG highlights how coupling part evidence with global shape hypotheses can improve manipulation when sensory cues are sparse.

Limitations and Future Work. First, CRAG is affected by distribution bias in training data. While PartNeXt (Wang et al., 2025a) is large, it overrepresents common categories with canonical part structures, leaving the long tail underrepresented and limiting OOD generalization, which motivates the need for broader community-scale part assembly data. Second, metrics like PA and CD measure geometry but can miss semantic correctness under symmetry and interchangeable parts, such as swapping identical table legs, calling for permutation and symmetry-aware evaluation. Finally, CRAG currently uses an image-conditioned interface inherited from TripoSG (Li et al., 2025b), whereas many applications need richer controls, such as sketches and language, for hypothesis-driven reconstruction and staged objectives. Overall, CRAG shows the promise of joint inference, but improving data coverage, evaluation, and multimodal controllability remains key future work.

Acknowledgment. This work was supported in part by NSF Grants 2152565, 2238968, and 2514030, and by NYU IT High Performance Computing resources, services, and staff expertise. This research was also supported by the NVIDIA Academic Grant Program using NVIDIA RTX 6000 Ada GPUs.

References

Abrams, G., Auguste, P., Pirson, S., De Groote, I., Halbrucker, É., Di Modica, K., Pironneau, C., Dedrie, T., Meloro, C., Fischer, V., et al. Earliest evidence of neanderthal multifunctional bone tool production from cave lion (panthera spelaea) remains. *Scientific Reports*, 15(1): 24010, 2025.

Boyer, D. M., Gunnell, G. F., Kaufman, S., and McGeary, T. M. Morphosource: archiving and sharing 3-d digital specimen data. *The Paleontological Society Papers*, 22: 157–181, 2016.

Chen, R., Zhang, J., Liang, Y., Luo, G., Li, W., Liu, J., Li, X., Long, X., Feng, J., and Tan, P. Dora: Sampling and benchmarking for 3d shape variational auto-encoders. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16251–16261, 2025.

Falcucci, A., Moroni, A., Negrino, F., Peresani, M., and Riel-Salvatore, J. The open aurignacian project: 3d scanning and the digital preservation of the italian paleolithic record. *Scientific Data*, 12(1):1037, 2025.

Ge, Y., Zhao, C., Wang, Y., and Wu, X. Robot-assisted autonomous reduction of a displaced pelvic fracture: a case report and brief literature review. *Journal of clinical medicine*, 11(6):1598, 2022.

Gower, J. C. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

Hahn, D. and Wojtan, C. High-resolution brittle fracture simulation with boundary elements. *ACM Transactions on Graphics*, 34(4):1–12, 2015.

Hahn, D. and Wojtan, C. Fast approximations for boundary element based brittle fracture simulation. *ACM Transactions on Graphics*, 35(4):1–11, 2016.

Hunyuan3D, T., Zhang, B., Guo, C., Liu, H., Yan, H., Shi, H., Huang, J., Yu, J., Li, K., Linus, Wang, P., Lin, Q., Liu, S., Yang, X., Tang, Y., Zhao, Y., Lai, Z., Liang, Z., and Zhao, Z. Hunyuan3d-omni: A unified framework for controllable generation of 3d assets. *arXiv preprint arXiv:2509.21245*, 2025.

Koutsoudis, A., Pavlidis, G., and Chamzas, C. Detecting shape similarities in 3d pottery repositories. In *2010 IEEE fourth international conference on semantic computing*, pp. 548–552, 2010.

Lee, N., Min, J., Lee, J., Kim, S., Lee, K., Park, J., and Cho, M. 3d geometric shape assembly via efficient point cloud matching. In *41st International Conference on Machine Learning*, pp. 26856–26873, 2024.

Lee, N., Min, J., Lee, J., Park, C., and Cho, M. Combinative matching for geometric shape assembly. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9540–9549, 2025.

Li, S., Jiang, Z., Chen, G., Xu, C., Tan, S., Wang, X., Fang, I., Zyskowski, K., McPherron, S. P., Iovita, R., Feng, C., and Zhang, J. Garf: Learning generalizable 3d reassembly for real-world fractures. In *International Conference on Computer Vision*, 2025a.

Li, Y., Zou, Z.-X., Liu, Z., Wang, D., Liang, Y., Yu, Z., Liu, X., Guo, Y.-C., Liang, D., Ouyang, W., et al. Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. *arXiv preprint arXiv:2502.06608*, 2025b.

Liu, R., Deng, K., Wang, Z., and Liu, C. Stablelego: Stability analysis of block stacking assembly. *IEEE Robotics and Automation Letters*, 9(11):9383–9390, 2024.

Lu, J., Sun, Y., and Huang, Q. Jigsaw: Learning to assemble multiple fractured objects. In *Advances in Neural Information Processing Systems*, volume 36, pp. 14969–14986, 2023.

Lu, J., Hua, G., and Huang, Q. Jigsaw++: Imagining complete shape priors for object reassembly. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6704–6714, 2025a.

Lu, J., Liang, Y., Han, H., Hua, J., Jiang, J., Li, X., and Huang, Q. A survey on computational solutions for reconstructing complete objects by reassembling their fractured parts. In *Computer Graphics Forum*, pp. e70081, 2025b.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Scarpellini, G., Fiorini, S., Giuliari, F., Moreiro, P., and Del Bue, A. Diffassemble: A unified graph-diffusion model for 2d and 3d reassembly. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28098–28108, 2024.

Sellán, S., Chen, Y.-C., Wu, Z., Garg, A., and Jacobson, A. Breaking bad: A dataset for geometric fracture and reassembly. In *Advances in Neural Information Processing Systems*, volume 35, pp. 38885–38898, 2022.

Sellán, S., Luong, J., Mattos Da Silva, L., Ramakrishnan, A., Yang, Y., and Jacobson, A. Breaking good: Fracture modes for realtime destruction. *ACM Transactions on Graphics*, 42(1):1–12, 2023.

Sun, T., Zhu, L., Huang, S., Song, S., and Armeni, I. Rectified point flow: Generic point cloud pose estimation. In *Advances in Neural Information Processing Systems*, 2025.

Wagner, M. (ed.). *Pottery and Chronology of the Early Iron Age in Central Asia*. The Kazimierz Michałowski Foundation and Institute of Archaeology, University of Warsaw, Warszawa, 2013.

Wang, P., He, Y., Lv, X., Zhou, Y., Xu, L., Yu, J., and Gu, J. Partnext: A next-generation dataset for fine-grained and hierarchical 3d part understanding. *arXiv preprint arXiv:2510.20155*, 2025a.

Wang, Z., Chen, J., and Furukawa, Y. Puzzlefusion++: Auto-agglomerative 3d fracture assembly by denoise and verify. In *International Conference on Learning Representations*, 2025b.

Wu, S., Lin, Y., Zhang, F., Zeng, Y., Xu, J., Torr, P., Cao, X., and Yao, Y. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. In *Advances in Neural Information Processing Systems*, volume 37, pp. 121859–121881, 2024a.

Wu, X., Jiang, L., Wang, P.-S., Liu, Z., Liu, X., Qiao, Y., Ouyang, W., He, T., and Zhao, H. Point transformer v3: Simpler, faster, stronger. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., and Yang, J. Structured 3d latents for scalable and versatile 3d generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21469–21480, 2025.

Zhang, B., Tang, J., Niessner, M., and Wonka, P. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Transactions On Graphics*, 42(4):1–16, 2023.

Zhao, C., Cao, Q., Sun, X., Wu, X., Zhu, G., and Wang, Y. Intelligent robot-assisted minimally invasive reduction system for reduction of unstable pelvic fractures. *Injury*, 54(2):604–614, 2023.

Zhao, W., Cao, Y.-P., Xu, J., Dong, Y., and Shan, Y. Assembler: Scalable 3d part assembly via anchor point diffusion. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, 2025a.

Zhao, Z., Lai, Z., Lin, Q., Zhao, Y., Liu, H., Yang, S., Feng, Y., Yang, M., Zhang, S., Yang, X., et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025b.