

# KINEDIFF3D: KINEMATIC-AWARE DIFFUSION FOR CATEGORY-LEVEL ARTICULATED OBJECT SHAPE RECONSTRUCTION AND GENERATION

## – SUPPLEMENTARY MATERIALS –

**Anonymous authors**

Paper under double-blind review

### 1 OVERVIEW

In this supplementary material, we first show a detailed related work in Sec. 2. Afterward, more methodology details are discussed in Sec. 3. In addition, we offer more information about our experiment in Sec. 4. Also, more qualitative results are shown in Sec. 5. A preview version of the core codes is attached for reproducibility in Sec. 6. Finally, we disclose the use of Large Language Models in Sec. 7.

### 2 RELATED WORK, EXTENDED

#### 2.1 POSE AND JOINT ESTIMATION FOR ARTICULATED OBJECTS, EXTENDED

Recent advances in articulated pose estimation include part-centric methods Li et al. (2020); Liu et al. (2022b) that independently optimize per-part poses but often violate kinematic constraints under occlusion, and generative approaches like Zhang et al. (2023) that leverage diffusion models for rigid objects yet lack explicit joint parameter modeling. Traditional pose estimation algorithms primarily focus on instance-level tasks Jin et al. (2025), where the pose of a specific object is estimated with the aid of its known CAD model. With advancements in computer vision and robotics, research has increasingly shifted toward category-level 6D pose estimation. This approach offers unique practical value by enabling pose generalization across unseen objects, allowing for reasonable 6D pose predictions even for novel instances.

KineDiff3D addresses this gap by integrating pose and joint estimation within a diffusion-based framework. By conditioning the diffusion model on partial point clouds, KineDiff3D estimates global pose (SE(3)) and joint parameters. Additionally, our Chamfer-distance-based optimization module refines these estimates during inference, improving accuracy for both synthetic and real-world data. This unified approach distinguishes KineDiff3D from prior methods that treat reconstruction and pose estimation as separate tasks.

### 3 METHODOLOGY, EXTENDED

#### 3.1 KINEMATIC-AWARE SHAPE PRIOR LEARNING, EXTENDED

As mentioned in the main paper, we regularize the latent space using KL divergence, and employ multi-task learning to predict SDF values, segmentation labels, and joint angles, ensuring information density and diversity. The training objective is defined as:

$$\mathcal{L}_{KA} = \lambda_1 \|SDF_Q - \hat{SDF}_Q\|_1 + \lambda_2 \mathcal{L}_{CE}(S, \hat{S}) + \lambda_3 \|A - \hat{A}\|_1 + \beta D_{KL}(\mathcal{N}(\mu, \sigma^2) \| \mathcal{N}(0, 0.25^2)) \quad (1)$$

The first term of the Eq 1 is the L1 loss between the predicted SDF values  $\hat{SDF}_Q \in \mathbb{R}^L$  for query points  $Q \in \mathbb{R}^{L \times 3}$  and their ground-truth values  $SDF_Q$ , ensuring accurate surface reconstruction. The second term is the cross-entropy loss between the predicted segmentation labels

$\hat{S} \in \{0, 1, \dots, K - 1\}^N$  and ground-truth labels  $S$ , promoting precise part segmentation. The third term is the L1 loss between the predicted joint angles  $\hat{A} \in \mathbb{R}^{K-1}$  and ground-truth angles  $A$ , ensuring accurate kinematic state reconstruction.

### 3.2 POSE AND JOINT ESTIMATION MODULE, EXTENDED

**Joint Estimation.** To maintain kinematic consistency throughout the articulated structure, we extend the diffusion-based prediction framework to joint parameters using the identical formulation applied to the base part pose estimation:

- **Revolute joints** are parameterized as 7D vectors:  $\mathbf{y}_r = [l_x, l_y, l_z, d_x, d_y, d_z, \theta_r]^\top$  combining 3D joint location  $l_r \in \mathbb{R}^3$ , 3D joint direction  $d_r \in \mathbb{R}^3$  (unit vector), and joint state  $\theta_r \in \mathbb{R}$
- **Prismatic joints** are compacted into 4D vectors:  $\mathbf{y}_p = [d_x, d_y, d_z, \theta_p]^\top$  containing 3D direction  $d_p \in \mathbb{R}^3$  and displacement  $\theta_p \in \mathbb{R}$

For articulated objects with  $K$  parts, all child joint parameters are aggregated into a single high-dimensional tensor:

$$y = \bigoplus_{k=2}^K y^{(k)} \quad (2)$$

where  $y^{(k)}$  corresponds to the joint connecting part  $k - 1$  to part  $k$ . Crucially,  $y$  undergoes the same conditional diffusion process as the base pose  $x$ , following the identical variance-exploding SDE formulation  $dy = g(t)dw$  described in main paper.

**Re-Scoring.** To improve the precision of articulation pose prediction, we employ an additional  $3N$ -channel multilayer perceptron (MLP) to compute confidence scores for rotation ( $\omega_R$ ), translation ( $\omega_t$ ), and joint parameters ( $\omega_J$ ). To address the challenge of noisy predictions in articulated objects, we generate  $M$  candidate poses to minimize the adverse effects of erroneous pose estimates on model performance. Specifically, to eliminate inaccurate predictions, we introduce a unified score,  $\mathcal{U}$ , defined as the harmonic mean of the confidence scores, which can be expressed as:

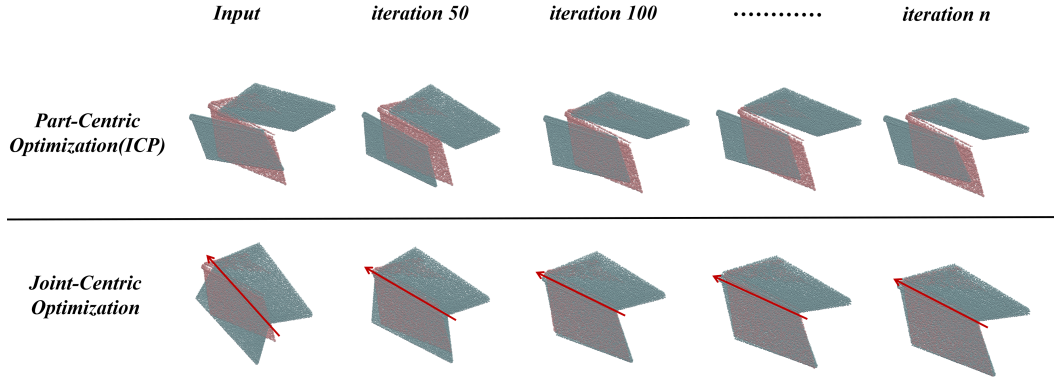
$$\mathcal{U} = \frac{3}{\frac{1}{\omega_R} + \frac{1}{\omega_t} + \frac{1}{\omega_J}} \quad (3)$$

Subsequently, the unified scores  $\mathcal{U}_{m=1}^M$  for the  $M$  candidates are ordered from highest to lowest. We then discard the bottom 40% of candidates and apply average pooling to the remaining candidates to determine the final predicted pose.

**Part-Segmentation.** To enable the Chamfer distance minimization outlined in Equation (9) of the main paper, which requires part-level segmentation of the input partial point cloud, we augment the diffusion-based pose estimation module with an additional segmentation head. This head outputs per-point part labels for the input  $O$ , allowing us to obtain the segmented point clouds  $O^{(k)}$  for each articulated part. This enhancement ensures accurate alignment during the iterative optimization phase while maintaining kinematic consistency.

characteristic	Part-Centric	Joint-Centric
Joint Param Optimization	×	✓
Kinematic constraints	×	✓
Global consistency	suboptimal	optimal
Sensitivity to initialization	high	low

Table 1: Comparison with Joint-Centric Optimization and Part-Centric Optimization.

Figure 1: **Part-centric Optimization(ICP) vs Joint-Centric Optimization.**

### 3.3 INFERENCE AND ITERATIVE OPTIMIZATION, EXTENDED

The KineDiff3D framework processes partial point clouds through a pipeline that co-optimizes poses, joints, and geometry. Our Joint-Centric Iterative Optimization (Figure 1 (bottom)) fundamentally diverges from Part-Centric approaches by enforcing kinematic integrity at every update cycle. While Part-Centric methods like ICP (Figure 1 (top)) treat parts as independent entities—leading to disjointed transformations and inconsistent final poses, our method synchronizes motion through explicit joint parameter optimization (✓ in Table 1). Crucially, every iteration propagates gradients along kinematic chains using Rodrigues rotation. This enforces deterministic pose hierarchies, eliminating Part-Centric’s tendency for broken articulations (× Kinematic constraints in Table 1).

As visualized in Figure 1 (bottom), our joint-centric loop achieves coherent global alignment in fewer iterations than ICP’s fragmented convergence (Figure 1, top). This efficiency stems from unified gradient propagation, adjusting base poses and joint parameters while reconstructing geometry conditioned on canonicalized inputs. Consequently, we attain optimal global consistency (Table 1) as errors cannot compound across independently optimized parts. By contrast, Part-Centric’s suboptimal consistency (Table 1) manifests as unstable Chamfer distances ( $L_{CD}$ ) under occlusion, where disconnected gradients cause drift. Detailed quantitative comparisons can be seen in (Ablation Study, Extend).

## 4 EXPERIMENTS, EXTENDED

### 4.1 EXPERIMENTS SETTINGS, EXTENDED

**Implementation Details.** We provide comprehensive implementation details for KineDiff3D: For training the Kinematic-Aware Shape Prior Module, the input consists of complete point clouds sampled at 1,024 points in canonical space with a batch size of 32; for the Pose and Joint Estimation Module, inputs are partial point clouds sampled at 1,024 points in camera space using a consistent batch size of 32; and for the Shape Reconstruction and Generation Module, inputs are derived by transforming these camera-space partial point clouds into canonical space using ground-truth poses, maintaining the same batch size of 32. All modules employ a cosine learning rate scheduler initialized at 0.001 and undergo 1,000 training epochs. Experiments are conducted on Ubuntu 22.04 using four NVIDIA RTX 4090 GPUs with 24GB memory, ensuring unified hardware configurations across all training and evaluation phases.

**End-to-End Training Strategy.** We employ a hybrid approach where the KA-VAE and dual diffusion models are first pre-trained separately to optimize efficiency, then fine-tuned end-to-end for enhanced geometric constraints and reduced error accumulation.

**Datasets.** We conducted our experiments using two datasets: the synthetic dataset ArtImage Xue et al. (2021) and the semi-synthetic dataset ReArtMix Liu et al. (2022a). For ArtImage, we adopted 223 articulated object instances across five categories (eyeglasses, scissors, laptop, dishwasher, and drawer). For ReArtMix, we utilized 48 instances across five categories (box, scissor, stapler, cutter,

and drawer). For reconstruction data processing, we implement a unified normalization and sampling pipeline as follows: We align all objects from the ArtImage dataset within a normalized object coordinate system (NOCS) Wang et al. (2019) by centering each object at the origin of a 3D cube with coordinates spanning  $(-1, -1, -1)$  to  $(1, 1, 1)$  and uniformly scaling them such that the diagonal of their tight bounding boxes equals 1. We sample 250,000 surface points per mesh to form a zero-distance point cloud, then generate two additional query points per surface point by adding Gaussian noise sampled from  $\mathcal{N}(0, 0.005)$  and  $\mathcal{N}(0, 0.0005)$  distributions. We supplement these with all points from a  $128 \times 128 \times 128$  grid spanning the normalized cube, computing signed distance values using ground-truth meshes. During training, each mesh iteration randomly selects 1,024 surface points for shape feature extraction and samples 16,000 query points (30% from the uniform grid, 70% from perturbed near-surface points) for SDF supervision, ensuring scale-invariant learning while capturing intricate geometric details through strategic near-surface sampling. Identical pre-processing procedures are applied to the ReArtMix dataset to maintain methodological consistency across all experimental data.

### Metrics.

**Reconstruction and Generation Task.** For evaluation metrics, we employ the Chamfer-L1 distance (CD) to measure reconstructed mesh quality: Chamfer-L1 Distance (CD) quantifies the discrepancy between reconstructed and ground-truth surfaces. For predicted point cloud  $P_{pred}$  and ground-truth  $P_{gt}$ , the bidirectional CD is computed as:

$$CD(P_{pred}, P_{gt}) = \frac{1}{2} \left( \frac{1}{|P_{pred}|} \sum_{m \in P_{pred}} \min_{n \in P_{gt}} \|m - n\| + \frac{1}{|P_{gt}|} \sum_{n \in P_{gt}} \min_{m \in P_{pred}} \|n - m\| \right) \quad (4)$$

$\|\cdot\|_1$  denotes Manhattan distance (L1 norm). 30,000 points are uniformly sampled per surface. Final CD values scaled by 1000. CD-w (Whole-object): Computed over all object parts. CD-s (Static parts): Evaluated only on non-articulated components (e.g., laptop base). Point sampling excludes movable regions. CD-m (Movable parts): Restricted to articulated components (e.g., laptop screen). Sampling focuses on joint-adjacent surfaces.

**Pose and Joint Estimation Task.** We evaluate kinematic properties using four complementary metrics:

- Rotation Error (Rot Err): Angular deviation ( $^\circ$ ) between predicted  $R_{pred}$  and ground-truth  $R_{gt}$  rotations:

$$\Delta R = \cos^{-1} \left( \frac{\text{tr}(R_{pred}^T R_{gt}) - 1}{2} \right) \times \frac{180}{\pi} \quad (5)$$

where  $\text{tr}(\cdot)$  is the matrix trace.

- Translation Error (Trans Err): Euclidean distance (m) between predicted  $t_{pred}$  and  $t_{gt}$ :

$$\Delta t = \|t_{pred} - t_{gt}\|_2 \quad (6)$$

- Joint State Error (error): Revolute joints: Absolute angle difference  $|\theta_{pred} - \theta_{gt}|$  ( $^\circ$ ) Prismatic joints: Absolute displacement difference  $|d_{pred} - d_{gt}|$  (m)
- Joint Direction Error (Ang Err): Angular deviation ( $^\circ$ ) of revolute joint axis:

$$\Delta d = \cos^{-1}(d_{pred} \cdot d_{gt}) \times \frac{180}{\pi} \quad (7)$$

- Joint Location Error (Pos Err): Euclidean distance (m) between joint centers  $l_{pred}$  and  $l_{gt}$ .

## 4.2 COMPARISON WITH THE SOTA METHODS, EXTENDED

**Pose and Joint Estimation Task.** We calculated the error of each part of the articulated objects on ArtImage with quantitative results detailed in Table 2. Compared to classical methods, we achieve the



Category	Method	Per-part Pose		Joint State	Joint Parameter	
		rotation error (°) ↓	translation error (m) ↓	error ↓	angle error (°) ↓	distance error (m) ↓
Laptop	A-NCSH Li et al. (2020)	5.3, 5.4	0.054, 0.043	3.5°	1.7	0.09
	Genpose Zhang et al. (2023)	5.3, 4.1	0.068, 0.060	3.4°	3.8	0.03
	ShapePose Zhou et al. (2025)	5.0, 4.6	0.052, 0.064	5.9°	3.3	0.06
	<b>KineDiff3D (Ours)</b>	<b>4.0, 3.7</b>	<b>0.034, 0.049</b>	<b>3.2°</b>	<b>0.9</b>	<b>0.03</b>
Eyeglasses	A-NCSH Li et al. (2020)	3.7, 22.3, 23.2	0.049, 0.313, 0.324	12.8°, 14.2°	3.1, 3.1	0.07, 0.06
	Genpose Zhang et al. (2023)	5.0, 7.4, 7.6	0.063, 0.113, 0.301	5.0°, 5.1°	4.1, 4.3	0.04, 0.05
	ShapePose Zhou et al. (2025)	4.2, 6.0, 6.0	0.049, 0.106, 0.108	5.7°, 5.6°	3.8, 3.9	0.05, 0.08
	<b>KineDiff3D (Ours)</b>	<b>3.1, 5.2, 5.4</b>	<b>0.047, 0.095, 0.088</b>	<b>4.9°, 4.9°</b>	<b>1.7, 1.8</b>	<b>0.03, 0.03</b>
Dishwasher	A-NCSH Li et al. (2020)	4.0, 4.8	0.059, 0.123	3.8°	6.1	0.11
	Genpose Zhang et al. (2023)	6.1, 6.3	0.115, 0.164	3.8°	4.8	0.09
	ShapePose Zhou et al. (2025)	3.9, 4.3	0.055, 0.079	6.0°	2.2	0.04
	<b>KineDiff3D (Ours)</b>	<b>2.9, 3.7</b>	<b>0.048, 0.055</b>	<b>2.3°</b>	<b>1.7</b>	<b>0.03</b>
Scissors	A-NCSH Li et al. (2020)	2.0, 2.6	0.035, 0.021	4.4°	0.8	0.04
	Genpose Zhang et al. (2023)	4.1, 3.5	0.050, 0.041	3.3°	2.8	0.06
	ShapePose Zhou et al. (2025)	2.3, 2.9	0.033, 0.045	4.2°	1.9	0.08
	<b>KineDiff3D (Ours)</b>	<b>2.1, 5.4</b>	<b>0.023, 0.021</b>	<b>2.5°</b>	<b>0.5</b>	<b>0.02</b>
Drawer	A-NCSH Li et al. (2020)	2.8, 3.5, 3.9, 2.9	0.045, 0.155, 0.157, 0.075	0.38m, 0.45m, 0.41m	2.6, 2.7, 5.2	-, -, -
	Genpose Zhang et al. (2023)	4.4, 4.4, 4.4, 4.4	0.111, 0.143, 0.144, 0.115	0.12m, 0.13m, 0.10m	3.3, 3.3, 3.3	-, -, -
	ShapePose Zhou et al. (2025)	3.2, 3.6, 3.5, 3.8	0.124, 0.178, 0.175, 0.121	0.62m, 0.78m, 0.68m	2.0, 2.3, 2.1	-, -, -
	<b>KineDiff3D (Ours)</b>	<b>2.8, 2.8, 2.8, 2.8</b>	<b>0.041, 0.085, 0.091, 0.071</b>	<b>0.52m, 0.66m, 0.52m</b>	<b>1.4, 1.8, 1.9</b>	<b>-, -, -</b>

Table 2: Comparison of pose and joint estimation with State-of-the-arts on ArtImage Dataset. The categories laptop, eyeglasses, dishwasher and scissors contain only free joint and revolute joints, and the drawer category contains free joint and prismatic joints.

best pose estimation results for the laptop category, with rotation error of 4.0°, 3.7°. In Dishwasher, the translation error is only 0.048m, 0.055m. Concerning joint state error, we achieve a remarkable 4.9°, 4.9° for category eyeglasses. This superiority directly validates our method’s effectiveness in integrating kinematic constraints during differentiable optimization. Qualitative results is provided in supplementary material.

Index	Occlusion Level (Visibility)	Reconstruction(CD-w)	Rot Err (°)	Trans Err (m)
I	Joint-centric	6.31	3.3	0.140
II	Part-centric	8.73	6.2	0.052

Table 3: **Ablation Study Results.** Note that experiments are conducted on the category Dishwasher.

Error Level	CD-w	CD-s	CD-m
±15° rot, 5cm trans	2.17	2.53	1.76
±30° rot, 10cm trans	3.14	3.62	2.84

Table 4: **Reconstruction Robustness Under Synthetic Pose Errors.** Note that experiments are conducted on the category Laptop.

### 4.3 ABLATION STUDY, EXTENDED

**Joint-Centric v.s Part-Centric.** Building on the joint-centric modeling for pose estimation, our iterative optimization module significantly enhances reconstruction accuracy by minimizing the bidirectional Chamfer distance ( $L_{CD}$ ) between the reconstructed mesh transformed into camera space and the input partial point cloud. This approach jointly refines the global pose and joint parameters through gradient descent, preserving kinematic dependencies that ensure part connectivity remains intact during updates, unlike part-centric methods that independently optimize each component, leading to misalignments and inflated errors. As show in Table 3 (I-II), for the dishwasher, joint-centric optimization achieves a CD-w of 6.31 (lower than part-centric’s 8.73) by ensuring door rotations align with hinge axes, while simultaneously reducing rotation errors to 3.3° (vs. 6.2°).

**Robustness to Initial Pose Estimation Errors.** KineDiff3D’s inference pipeline demonstrates significant resilience to large initial pose errors through integrated architectural safeguards and iterative refinement. While substantial miscalibration of the initial SE(3) pose could theoretically distort the canonical transformation of partial point clouds, potentially propagating errors through the conditional diffusion process, our framework mitigates this via three synergistic mechanisms: First, the Latent diffusion model inherently accommodates spatial perturbations through its cross-attention

mechanism between noisy latent codes and locally invariant geometric features extracted by the PointNet++ encoder  $\Gamma$ , enabling reconstruction fidelity even under moderate canonical space misalignment. Second, the joint-centric optimization loop actively compensates for initial errors by bidirectionally minimizing Chamfer distance through gradient-based refinement of both global pose and joint parameters while strictly preserving kinematic constraints. Third, empirical stress testing confirms graceful degradation rather than catastrophic failure. To quantify KineDiff3D’s tolerance to inaccuracies in the initial pose estimation phase, we conduct controlled experiments injecting synthetic rotation/translation errors into the initial pose estimates. As shown in Table 4, reconstruction quality degrades linearly rather than catastrophically under extreme perturbations. This demonstrates our framework’s resilience to real-world sensor noise and pose ambiguity.

## 5 QUALITATIVE RESULTS, EXTENDED

We provide comprehensive qualitative evaluations across both reconstruction/generation and pose estimation tasks. For reconstruction and generation, results on the synthetic ArtImage dataset are visualized in Figure 2, while semi-synthetic ReArtMix demonstrations appear in Figure 3. For articulated pose estimation, qualitative analyses on ArtImage and ReArtMix objects are presented in Figure 4 and 5 respectively, covering diverse articulation states and object categories.

## 6 REPRODUCIBILITY, EXTENDED

In order to ensure the reproducibility of our method, we provide the codes of core modules and instructions, Please click **codes-path**(cd: ./codes) for codes.

## 7 LARGE LANGUAGE MODELS (LLMs) USAGE DISCLOSURE, EXTENDED

**Purpose and Scope of Use:** LLMs solely served as a general-purpose writing and editing aid. Their role was strictly limited to performing language polishing, grammar checking, wording refinement, and improving the fluency of texts originally drafted by the human authors. LLMs did not generate any original ideas, data, or conclusions.

**Dominance of Human Authorship:** All core academic content—including, but not limited to, research ideation, theoretical framing, methodology design, experimentation and analysis, result interpretation, and conclusion drawing—was independently conceived and completed by the human authors. At no point did LLMs participate in any core academic decision-making processes that required scientific insight or domain expertise.

**Tools and Prompts:** We utilized LLM tools such as Deepseek and Claude. The instructions (prompts) provided to the LLMs were inherently non-creative and primarily limited to directives such as “improve the grammar of this paragraph,” “make this academic expression more concise,” “suggest more professional synonyms for this sentence,” or “check for terminological consistency.”

**Accountability:** The human authors conducted rigorous review, verification, and final approval of every part of the manuscript. We assume full responsibility for all information, claims, and viewpoints presented in the paper.

## REFERENCES

- Le Jin, Guoshun Zhou, Zherong Liu, Yuanchao Yu, Teng Zhang, Minghui Yang, and Jun Zhou. Irpe: Instance-level reconstruction-based 6d pose estimator. *Image and Vision Computing*, 154: 105340, 2025.
- Xiaolong Li, He Wang, Li Yi, Leonidas Guibas, A. Lynn Abbott, and Shuran Song. Category-level articulated object pose estimation, 2020. URL <https://arxiv.org/abs/1912.11913>.
- Liu Liu, Wenqiang Xu, Haoyuan Fu, Sucheng Qian, Yang Han, and Cewu Lu. Akb-48: A real-world articulated object knowledge base, 2022a. URL <https://arxiv.org/abs/2202.08432>.

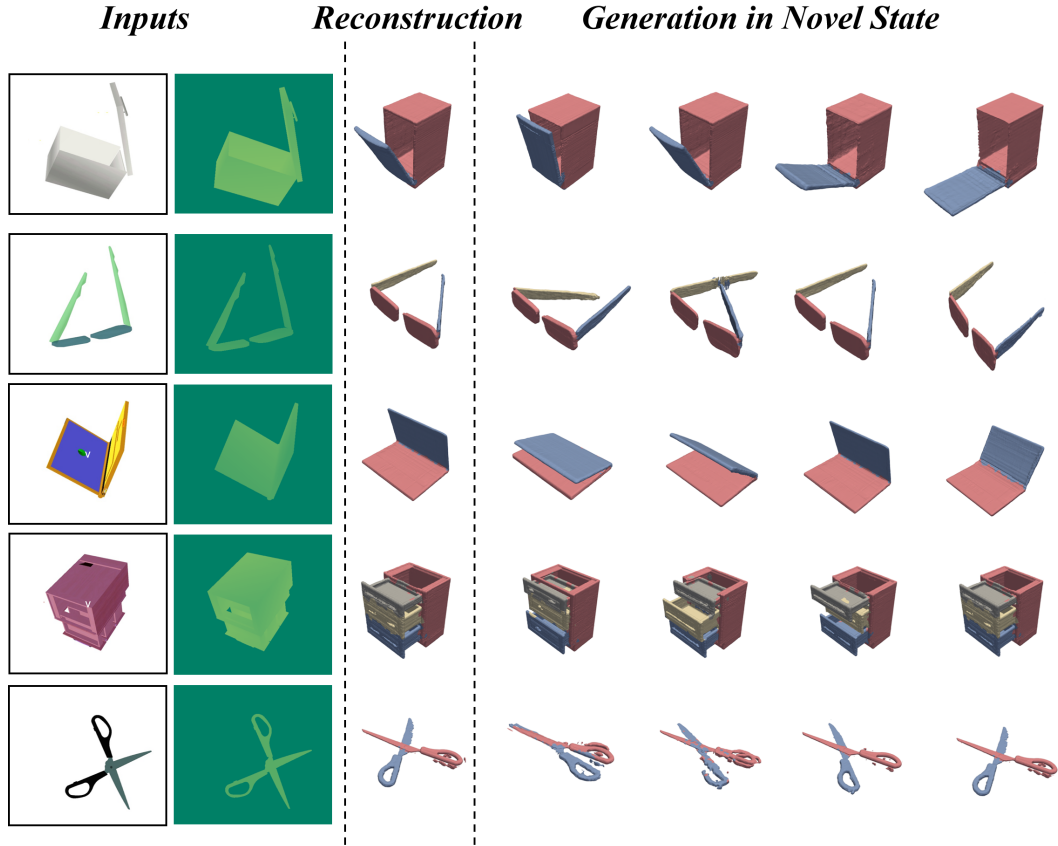


Figure 2: Qualitative Reconstruction and Generation Results on the ArtImage Datasets.

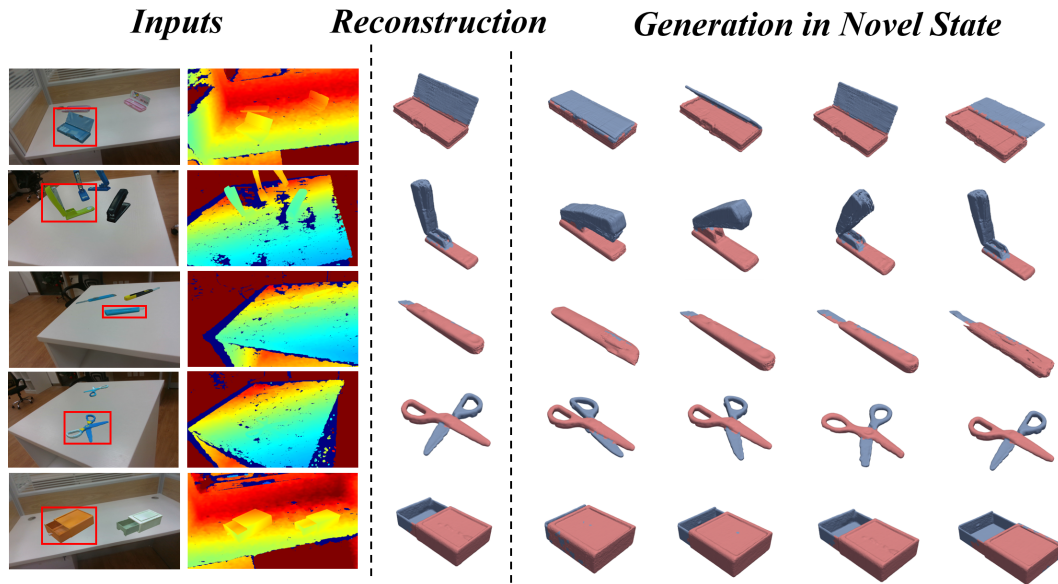


Figure 3: Qualitative Reconstruction and Generation Results on the ReArtMix Datasets.

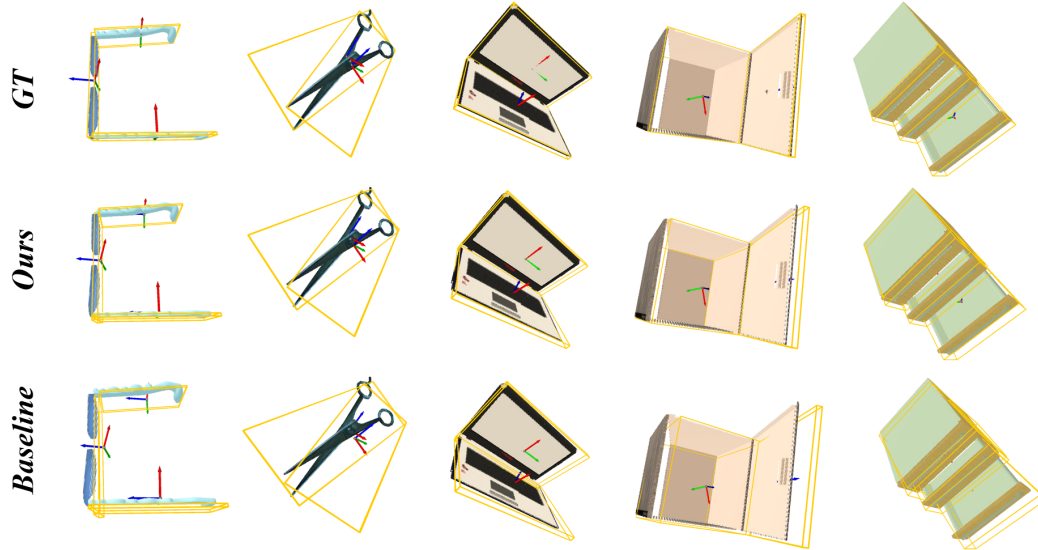


Figure 4: Qualitative Pose Estimation Results on the ArtImage Datasets.

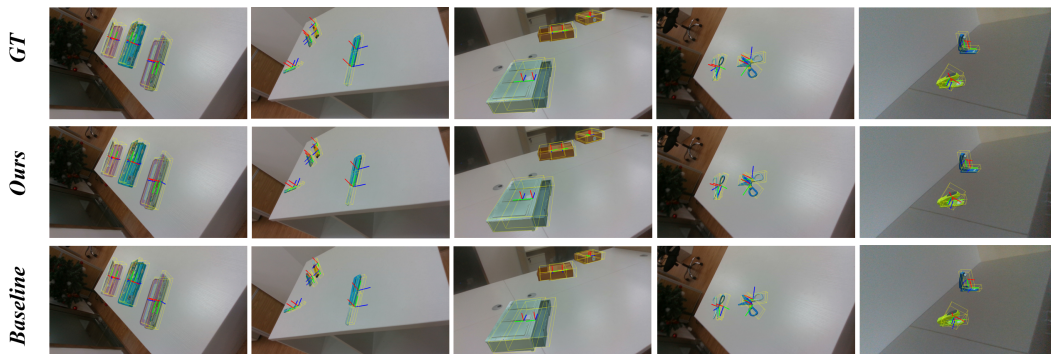


Figure 5: Qualitative Pose Estimation Results on the ReArtMix Datasets.

- Liu Liu, Han Xue, Wenqiang Xu, Haoyuan Fu, and Cewu Lu. Toward real-world category-level articulation pose estimation. *IEEE Transactions on Image Processing*, 31:1072–1083, 2022b.
- He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J. Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation, 2019. URL <https://arxiv.org/abs/1901.02970>.
- Han Xue, Liu Liu, Wenqiang Xu, Haoyuan Fu, and Cewu Lu. Omad: Object model with articulated deformations for pose estimation and retrieval, 2021. URL <https://arxiv.org/abs/2112.07334>.
- Jiyao Zhang, Mingdong Wu, and Hao Dong. Genpose: Generative category-level object pose estimation via diffusion models. *arXiv preprint arXiv:2306.10531*, 2023.
- Jun Zhou, Kai Chen, Mingqiang Wei, Xiao-Ping Zhang, Qi Dou, and Jing Qin. Canonical shape reconstruction with  $se(3)$  equivariance learning for weakly-supervised object pose estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2025. doi: 10.1109/TCSVT.2025.3542089.