PhysPart: Physically Plausible Part Completion for Interactable Objects

Rundong Luo*[1], Haoran Geng*[2,3], Congyue Deng[3], Puhao Li[4], Zan Wang[4], Baoxiong Jia[4], Leonidas Guibas[3], and Siyuan Huang[4]

[1] Cornell University
[2] University of California, Berkeley
[3] Stanford University
[4] Beijing Institute for General Artificial Intelligence (BIGAI)
* Equal Contribution

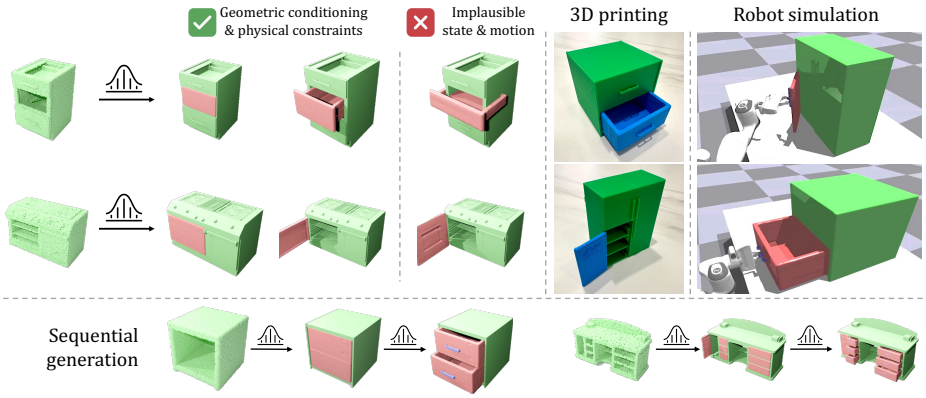

Fig. 1: Left: Given an incomplete object base, we propose a diffusion-based model to generate physically plausible parts that allow articulated interactions. **Right:** Our model naturally applies to real-world scenarios and embodied AI tasks, including 3D printing and robot manipulation, where physical plausibility is essential to the object interactivity and success of the tasks. For example, a small surface bump or a slight size mismatch can get the part stuck during interactions. **Bottom:** Our model supports sequential generation of dependent parts (*e.g.,* cabinet→door→handle) for more complex object hierarchies.

Abstract. Interactable objects are ubiquitous in our daily lives. Recent advances in 3D generative models make it possible to automate the modeling of these objects, benefiting a range of applications from 3D printing to the creation of robot simulation environments. However, while significant progress has been made in modeling 3D shapes and appearances, modeling object physics, particularly for interactable objects, remains challenging due to the physical constraints imposed by inter-part motions. In this paper, we tackle the problem of physically plausible

part completion for interactable objects, aiming to generate 3D parts that not only fit precisely into the object but also allow smooth part motions. To this end, we propose a diffusion-based part generation model that utilizes geometric conditioning through classifier-free guidance and formulates physical constraints as a set of stability and mobility losses to guide the sampling process. Additionally, we demonstrate the generation of dependent parts, paving the way toward sequential part generation for objects with complex part-whole hierarchies. Experimentally, we introduce a new metric for measuring physical plausibility based on motion success rates. Our model outperforms existing baselines over shape and physical metrics, especially those that do not adequately model physical constraints. We also demonstrate our applications in 3D printing, robot manipulation, and sequential part generation, showing our strength in realistic tasks with the demand for high physical plausibility.

1 Introduction

Creating 3D objects that facilitate human interactions is a critical and long-standing problem in industrial engineering, with applications ranging from manufacturing to developing simulation environments. With the recent proliferation of AI-facilitated 3D content creation [11, 34, 38–40, 51, 54, 71, 84], a line of efforts has been made to model interactable objects, particularly articulated ones, with 3D generative models [29, 37]. Imagine having a cabinet with a broken drawer; rather than engaging in labor-intensive carpentry with jigs and saws, a more efficient solution would be to generate a 3D model of the drawer and print it using a 3D printer, thus eliminating the need for specialized human expertise.

However, unlike rigid objects, whose modeling primarily focuses on shapes and appearances, interactable object modeling presents the additional challenge of intricate part motions under human interactions, constrained by geometry and physics. This complexity is particularly problematic for real-world applications, where high accuracy is essential. For instance, in the case of a drawer, even a slight size mismatch or a minor surface bump can cause it to get stuck in the cabinet.

In this paper, we address the problem of generating 3D object parts that are physically plausible both in their rest states and under interactions. Previous work on physics-constrained shape generation has predominantly concentrated on rigid objects, emphasizing connectivity [22, 23, 43] or structural stability [42]. However, a challenge arises for interactable objects with part motions: while the generated shapes are static, the inherent motion state is also crucial for physical plausibility. Thus, constraints on both rest and motion states must be carefully considered.

We propose a diffusion-based framework for physically plausible part completion for interactable objects. Given an object point cloud with a missing part, we represent the part as an implicit surface and generate its shape with a latent diffusion model on quantized latent embeddings, To guide the diffusion sampling process toward a physically plausible solution, we introduce two types

of conditions: the *geometric conditioning* applied via classifier-free guidance and the *physical constraints* formulated as inference-time losses applied to the score functions. Specifically, we design two sets of losses for physical constraints: stability losses that consider collisions and contacts at rest states and mobility losses that consider part motions. To incorporate the physical losses into the sampling process, we approximate their gradients using a prediction of the noise-free sample at each time step and add the approximated gradients to the score function. Finally, we also demonstrate the generation of dependent parts, paving the way toward sequential part generation for objects with complex part-whole hierarchies.

For evaluation, we construct a benchmark with metrics for assessing both the shapes and the physics of the generation results. Besides the widely adopted metrics for 3D generative models, we introduce a *physical plausibility metric* that directly evaluates the success rates of part motions. With both qualitative and quantitative results, we show that our framework generates shapes of high quality, outperforming methods that lack proper modeling of geometric conditioning or physical constraints. Furthermore, we demonstrate two downstream applications: 3D printing and robot manipulation, showcasing our strong potential in real-world applications that require high physical plausibility.

To summarize, our key contributions are:

- We introduce the task of physically plausible part completion for interactable objects and provide a benchmark with holistic evaluation metrics for both shape and physics.
- We propose a novel diffusion-based part generative model; it incorporates geometric conditioning and physical constraints by formulating them as stability and mobility losses, respectively.
- Experimentally, we show compelling results in both shape quality and physical plausibility compared with baseline models, demonstrating the efficacy of our model.
- We showcase physics-demanding applications of our model, including 3D printing, robot manipulation, and sequential part generation, where high physical plausibility is crucial to the task.

2 Related Work

Articulated Object Modeling. Articulated object modeling is a crucial and longstanding field in 3D vision and robotics, encompassing a wide range of work in perception [12,15–19,31,35,74], reconstruction [2,25,46,67], and generation [29,37]. These contributions have greatly advanced applications in simulation [19, 73] and robot manipulation [4, 13, 15, 16, 30, 32, 68, 72, 75]. Beyond these perspectives, the intricate per-part state changes and inter-part motions, which formulate how objects interact and respond to interactions, also play important roles. For instance, [17] categorizes parts based on their actions, [74] learns part segmentation by analyzing point clouds during part motions, and [25, 46, 67] focus on reconstructing parts and joints across various motion states. However,

most studies primarily focus on part motions at the pose level, often neglecting the critical interplay of underlying physics with detailed part geometry. This oversight is particularly significant in 3D generation, where physical plausibility is essential for the practical application of generated models. Nonetheless, this critical aspect of modeling remains under-explored in recent studies, which focus more on structure and shape [29, 37]. We direct readers to survey [36] for a more comprehensive overview.

Physics-Aware Shape Generation. Previous research in physics-aware 3D object generation has primarily focused on rigid shapes without complex self-interactions [22, 23, 42, 43]. For instance, [22, 23, 43] have used persistence diagrams to ensure topological connectivity, and [42] introduced a differentiable physical simulation layer to enhance stability under gravity. In contrast, physical principles have been extensively applied in other areas, such as in PhyScene [73], which aids in synthesizing interactable 3D scenes, and in human pose and motion generation where studies [69, 80, 81] incorporate physics-aware constraints to reduce collisions and enhance realism. Additionally, built on diffusion models, SceneDiffuser [24] and PhysDiff [77] achieve physically plausible generation by integrating physical optimization guidance into the denoising process. Our work uniquely extends these principles by embedding physical constraints, including stability and mobility, into a generative diffusion model, enabling the generation of physically plausible parts for interactable articulated objects.

3D Shape Completion. Shape completion is a vital step for reconstructing the missing parts of 3D shapes, traditionally tackled with techniques such as Laplacian hole filling [48, 61, 82] and Poisson surface reconstruction [26, 27]. These methods primarily address small gaps and basic geometric forms. Another strategy employs structural regularities, like symmetries, to infer unobserved parts of shapes [44, 50, 55, 62, 66]. With the advent of extensive 3D datasets, retrieval-based approaches [28, 33, 47, 64] have emerged, searching databases to find the best matches for incomplete inputs, alongside learning-based methods [6, 8, 10, 14, 20, 49, 60, 76] that minimize discrepancies between network predictions and actual shapes. Notably, 3D-EPN [10] utilizes a 3D encoder-decoder architecture to predict complete shapes from partial volumetric data, while Scan2Mesh [9] transforms range scans into 3D meshes through direct mesh surface optimization. PatchComplete [53] leverages local structural priors to fill in shapes from unseen categories. Alternatively, generative methods like GANs [3, 56, 70, 79, 83] and AutoEncoders [1, 45] offer a different approach by generating diverse, plausible shapes from partial inputs, though they often compromise on completion accuracy. In contrast, our method leverages diffusion models and introduces physics-aware loss guidance, effectively reducing surface artifacts and surpassing current state-of-the-art methods in producing realistic and physics-plausible 3D shapes, allowing applications like 3D printing and robot manipulation.

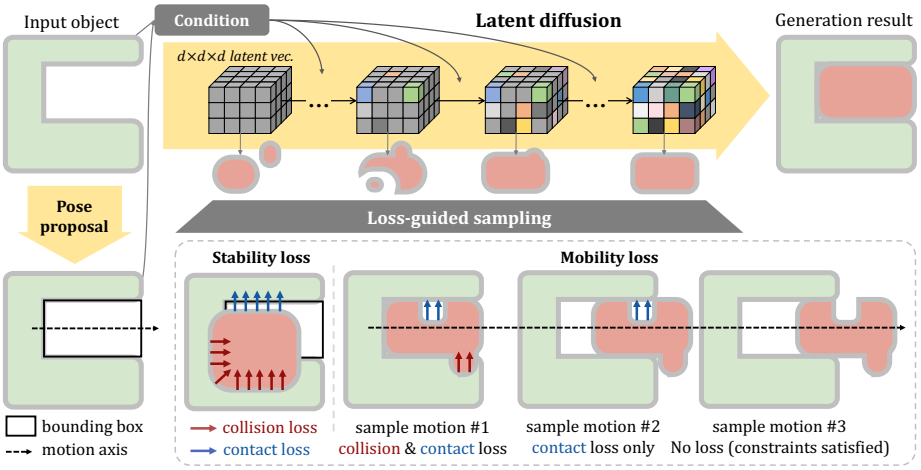

Fig. 2: Pipeline of our proposed framework. We train a pose proposal model to predict the missing part's bounding box and a latent diffusion model conditioned on the input object's point cloud and the missing part's bounding box within the latent space of part SDF. During inference, the trained pose proposal model first predicts the missing part's bounding box. We then apply the proposed physical-aware losses (contact and collision losses in static or dynamics states) to guide the sampling process.

3 Method

Given the full point cloud of an articulated object $\mathcal{O}$ with a missing part of a known category (*e.g.,* drawer), we aim to generate a replacement part p that integrates seamlessly into the object while adhering to geometric and physical constraints. We employ volumetric Signed Distance Fields (SDF) to model object parts and train a 3D-VQVAE to compress these models into a discrete and compact latent space. Unlike prior research that employs SDF representation overlooks the object's scale as generating SDF usually involves normalizing the part to the unit cube, our framework also emphasizes accurately scaling the objects for manufacturing purposes. Upon this part representation, our method utilizes a coarse-to-fine approach, involving two diffusion-based sub-models: a pose proposal model that determines the part's approximate position at the coarse level, and a part code generator that refines the shape based on geometric conditions (Sec. 3.1) and physical constraints (Sec. 3.2). This integrated framework ensures that the generated part can interact properly with external forces and fit accurately within the existing object structure. An overview of our pipeline is shown in Fig. 2.

Part-Pose Proposal. We propose a diffusion-based pose estimator to locate the bounding box of the part to be generated given the object point cloud $\mathcal{P}$. Inspired by GenPose [78], we train a score-based diffusion model Φ_θ and an energy-based diffusion model Ψ_ϕ. Then we first generate pose candidates $\{p^i\}_{i=1}^K$ from Φ_θ and then compute the pose energies $\Psi_\phi(p^i, \mathcal{P})$ for candidates via Ψ_ϕ.

After that, by ranking the candidates with the energies and then filtering out low-ranking candidates, the remaining bounding box candidates are aggregated into the final output using mean pooling.

3.1 Geometry-Conditioned Part Generation

SDF Generation with Latent Diffusion. SDFusion [5] introduces a diffusion-based 3D generative model. The approach begins by compressing high-dimensional SDFs into a compact latent space using a 3D-VQVAE. Specifically, an SDF $X \in \mathbb{R}^{D\times D\times D}$ is encoded into a latent vector $z \in \mathbb{R}^{d\times d\times d}$ by the encoder E_ϕ. The encoded SDF is then reconstructed by the decoder D_τ after passing z through a vector quantization codebook VQ, *i.e.,*

$$z = E_\phi(X), \text{ and } X' = D_\tau(\mathrm{VQ}(z)). \tag{1}$$

Afterward, a conditional diffusion model is trained on the learned latent distribution. Techniques including the classifier-free guidance [21] and DDIM [58] sampling are utilized to enhance the flexibility and efficiency of the generation process.

Geometric Conditioning. The above framework allows the generation of missing parts using point cloud as the condition. However, due to limited training data and the point cloud's inherent ambiguity, relying solely on the point cloud often yields unsatisfactory results. Thus, we leverage the part bounding box predicted by our pose estimation model as an additional condition. Specifically, given the predicted bounding box of the desired part, we use the ratio of the dimensions along the three axes as an auxiliary condition for the diffusion model. The length ratio rather than absolute values is used since SDF representation is scale-agnostic. Note that although we allow rotation to the input point cloud, the mesh is always generated in its canonical pose.

Empirically, we adopt the PointNet++ [52] (ϕ_1) to encode the point cloud $c_1 \in \mathbb{R}^{N\times 3}$, and a sinusoidal positional encoding followed by a two-layer MLP (ϕ_2) to encode the bounding box ratio $c_2 \in \mathbb{R}^3$. Then, we concatenate two control signals then inject them into the model using self-attention layers. The training objective for the diffusion model θ and encoders ϕ_1, ϕ_2 is formulated as:

$$L(\theta, \phi_1, \phi_2) := \mathbb{E}_{z,c,e,t}\left[\|\epsilon - \epsilon_\theta(z_t, t, D(\phi_1(c_1)), D(\phi_2(c_2)))\|^2\right], \tag{2}$$

where $D(\cdot)$ denotes a dropout operation, c_1, c_2 are the conditions, and z_t is the noisy latent variable at timestamp t. Please refer to Appendix A for details on model architecture.

Finally, using the predicted bounding box from the pose proposal model, we scale the mesh by the cube root of the volume ratio between the predicted bounding box and the axis-aligned bounding box of the generated mesh. This ensures the mesh is sized appropriately to fit the object's empty region.

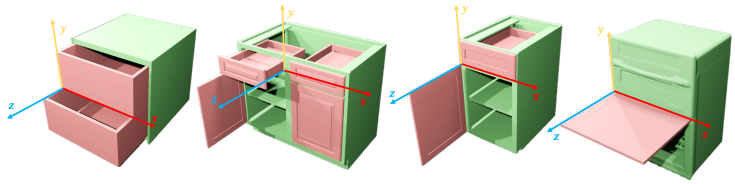

Fig. 3: Normalized 3D coordinate and shared motion constraints for certain part categories. Specifically, drawers could be pulled out along the $+z$ axis, while doors could rotate around $+x$, $-x$, $+y$, and $-y$ axes.

3.2 Loss-Guided Physical Constraints

Given an object's point cloud, we predict the bounding box for the missing part, generate the desired part by sampling from the diffusion model using the object's point cloud and the bounding box as conditions, and scale the generated mesh using the bounding box volume ratio. However, this generation process is agnostic to physical constraints and may result in parts incompatible with the original object. To address this, we introduce physics-aware losses to guide the diffusion sampling process, ensuring physical compatibility.

Stability Losses. The most critical criterion for the generated part is to avoid collision with the provided object point cloud $\mathcal{P}$. Therefore, we define the collision loss by considering the predicted part-SDF value $X(p)$ of the object point p as follows:

$$\ell_{collision}(X, \mathcal{P}) = \sum_{p \in \mathcal{P}} \text{ReLU}(-X(p) - \alpha), \tag{3}$$

where α represents a margin of tolerance. This loss function penalizes the SDF values when object points are inside the predicted SDF (*i.e.,* $X(p) < 0$), ensuring no parts of the object intersect the generated part. Additionally, maintaining proximity between the generated part and the object is crucial. Thus, we define contact loss to ensure no part of the object is beyond a distance α from the generated part:

$$\ell_{contact}(X, \mathcal{P}) = \min_{p \in \mathcal{P}} \text{ReLU}(X(p) - \alpha), \tag{4}$$

which penalizes excessive distances between the object points and the predicted SDF. Given that all parts are generated in a scale-agnostic canonical space, we adjust the scale of each part to match the predicted volume and apply transformations according to the predicted pose during loss calculations.

Mobility Losses. As shown in Fig. 3, certain part categories must follow shared motion constraints. For instance, drawers can move "out," and hinge doors can rotate around one of the four potential rotation axes. Thus, we extend both stability losses with motion. Since the joints of these parts have only one degree of freedom, we could ensure the feasibility of consistent part motion by random

sampling the part states. The enhanced loss functions are defined as follows:

$$\ell_{collision-m}(X, \mathcal{P}) = \mathbb{E}_{\hat{X} \sim M(X)} \sum_{p \in \mathcal{P}} \mathrm{ReLU}(-\hat{X}(p) - \alpha), \tag{5}$$

$$\ell_{contact-m}(X, \mathcal{P}) = \mathbb{E}_{\hat{X} \sim M(X)} \min_{p \in \mathcal{P}} \mathrm{ReLU}(\hat{X}(p) - \alpha). \tag{6}$$

where $M(X)$ is a set of transformed SDF. Note that for hinge doors, the ground truth rotation axis is unavailable during inference. However, as all four potential axes impose similar constraints when rotating from 0 to 90 degrees, we opt for a shared axis for all hinge door instances. To ensure both losses are differentiable, we opt to transform the point cloud instead of transforming the SDF directly.

Loss-Guided Sampling. Conditional diffusion models aim to generate samples x_0 conditioned on a specific criterion y, where the conditional score at timestamp t is derived using Bayes' rule:

$$\begin{aligned} \nabla_{x_t} \log p_t(x_t|y) &= \nabla_{x_t} \log p_t(x_t) + \nabla_{x_t} \log p_t(y|x_t) \\ &= -\frac{1}{\sqrt{1-\overline{\alpha}_t}} \epsilon_\theta(x_t, t) + \nabla_{x_t} \log p_t(y|x_t) \\ &= -\frac{1}{\sqrt{1-\overline{\alpha}_t}} (\epsilon_\theta(x_t, t) - \sqrt{1-\overline{\alpha}_t} \cdot \nabla_{x_t} \log p_t(y|x_t)). \end{aligned} \tag{7}$$

Hence, the ω-weighted modified score $\overline{\epsilon}_\theta$ is expressed as:

$$\overline{\epsilon}_\theta(x_t, t, y) = \epsilon_\theta(x_t, t) - \sqrt{1-\overline{\alpha}_t}\, \omega \nabla_{x_t} \log p_t(y|x_t), \tag{8}$$

where the gradient term $\nabla_{x_t} \log p_t(y|x_t)$ is written as $-\frac{1}{\sqrt{1-\overline{\alpha}_t}}(\epsilon_\theta(x_t, t, y) - \epsilon_\theta(x_t, t))$ in classifier-free guidance [21].

To introduce physical constraints into the sampling process, we approximate the gradient using our defined physics-aware loss functions [7,59]. As loss functions are defined on noise-free samples, we compute the loss on the predicted noise-free sample at timestamp t, denoted as $\hat{x}_0$:

$$\begin{aligned} \nabla_{x_t} \log p_t(y|x_t) &\approx \nabla_{x_t} \log p_t(y|\hat{x}_0) && \text{(Eq. (15) in [7])} \\ &= \nabla_{x_t} \log \frac{\exp(-\ell_y(\hat{x}_0))}{Z} = -\ell_y(\hat{x}_0). \end{aligned} \tag{9}$$

Finally, we combine classifier-free guidance and loss guidance to obtain the modified score:

$$\overline{\epsilon}_\theta(x_t, t, y) = \epsilon_\theta(x_t, t) + \omega_1(\epsilon_\theta(x_t, t, y) - \epsilon_\theta(x_t, t)) + \omega_2 \ell_y(\hat{x}_0), \tag{10}$$

where the ω_1, ω_2 are guidance weights. Denote X_0 to be the decoded SDF of latent code x_0, the $\ell_y(x_0)$ is the sum of our defined physical-aware losses:

$$\ell_y(x_0) = \ell_{collision} + \ell_{contact} + \ell_{collision-m} + \ell_{contact-m}. \tag{11}$$

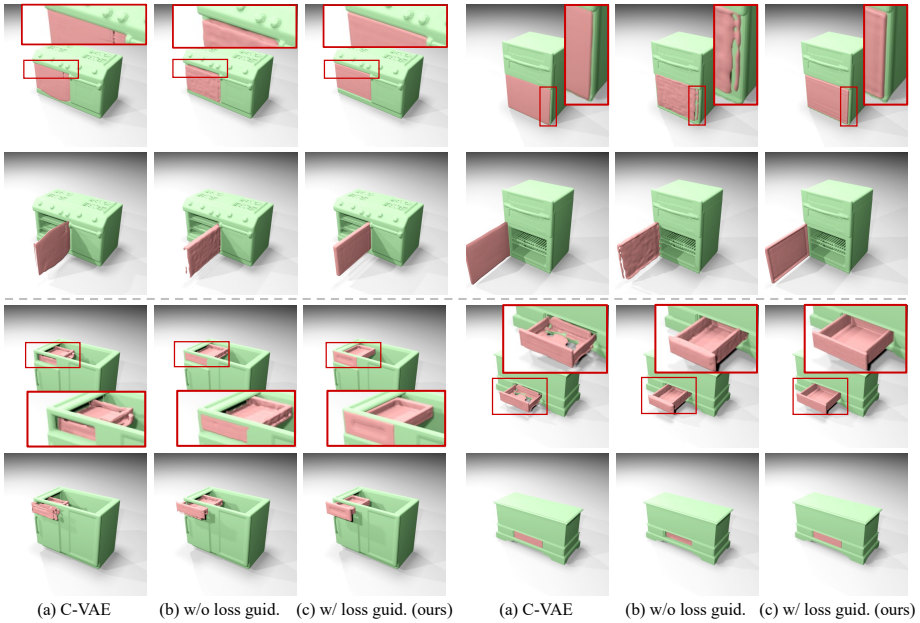

Fig. 4: Results of self-moving part generation. All visualizations are done in the objects' canonical poses, with the base shapes in green and the generated parts in red. While all methods can generate part shapes that are *roughly* reasonable, the bumpy surfaces and size mismatches in baseline results hinder their *physical plausibility*.

4 Experiments

This section demonstrates the effectiveness of our method in generating four types of common articulated objects: slider drawers, hinge doors, hinge knobs, and line handles, where the first two allow *self-movements* and the latter two must be attached to other parts (denoted as *dependent* parts). We then explore its potential applications in 3D printing, robot manipulation, and sequential part generation tasks. Appendix B presents more experimental results.

Data and Implementation. We train our models on the GAPartNet dataset [17], a part-centric object model dataset with part-level annotations and actionability alignment across part classes. Approximately 90% of instances are allocated for training for each part category, with the remainder reserved for evaluation. All part categories utilize the same VQVAE model, trained across all part categories at an input SDF resolution of 128, while a separate diffusion model is tailored for each category. To ensure comparability, both our method and all baselines involve random rotation of the input point clouds around the vertical axis during training. For testing, we sample three different rotation angles for each object instance. Further implementation details are available in Appendix A.

Baselines. Our primary baseline is the Conditional 3D Variational Autoencoder (3D-CVAE) [57], which learns a VAE on part SDF with the object's point cloud

Table 1: Quantitative generation results on self-moving parts. "Phy." denotes our physical plausibility metric.

Method	Slider drawer			Hinge door		
	CD ↓	F-score ↑	Phy. (%)	CD ↓	F-score ↑	Phy. (%)
3D-CVAE	0.00248	0.683	38.9	0.00106	0.867	45.6
Ours (w/o loss guidance)	0.00082	0.818	49.3	0.00083	0.873	71.1
Ours (w/ loss guidance)	**0.00060**	**0.862**	**74.3**	**0.00081**	**0.883**	**92.2**

as the condition. To ensure fair comparisons, we scale the generated parts using our predicted pose and then use the same settings to process the results. To further demonstrate the effectiveness of our loss guidance, we also compare the vanilla SDF diffusion without our proposed loss functions.

4.1 Physics-Plausible Generation

Metrics. We introduce a *physical plausibility* metric to evaluate the viability of generated shapes. Specifically, we first pre-process the meshes, excluding those with multiple large connected parts or those not watertight as physically infeasible. For the remaining meshes, a grid search around the part's ground truth position is conducted to locate positions where both collision loss (Eq. (5)) and contact loss (Eq. (6)) fall below a specified threshold. A part is deemed physically plausible if its movement along the $+x$ and $-x$ axes (for slider drawers) or along the $-z$ axis (for other part categories) causes these losses to exceed the threshold. Notably, for hinge doors, a shared rotation axis is used during loss-guided sampling, while the ground truth rotation axis is used for evaluation. Please see Appendix A.1 for a comprehensive explanation and pseudo-code. Besides the physical plausibility metric, we also utilize Chamfer Distance (CD) and F-score (@2%) [65] for evaluation.

Self-moving Part Generation. As shown in Tab. 1, our method significantly outperforms the baseline in terms of generation quality (CD and F-score) and physical plausibility in generating self-moving parts. The results also highlight the essential role of our proposed physical guidance in the generation process. Notably, as illustrated in Fig. 4, while the baseline method or the absence of physical guidance often leads to collisions or misalignments at the interface between the object and the generated part, our approach consistently delivers parts that seamlessly integrate with the original object, establishing a robust foundation for subsequent dependent part generation.

Dependent Part Generation. As shown in Tab. 2, our approach remains superior to baseline methods dependent parts. Fig. 5 reveals that while the baseline method frequently produces a degenerated distribution (such as generating oval-shaped handles regardless of input or producing broken knobs), our method achieves consistent high-fidelity results that meet the physical requirements.

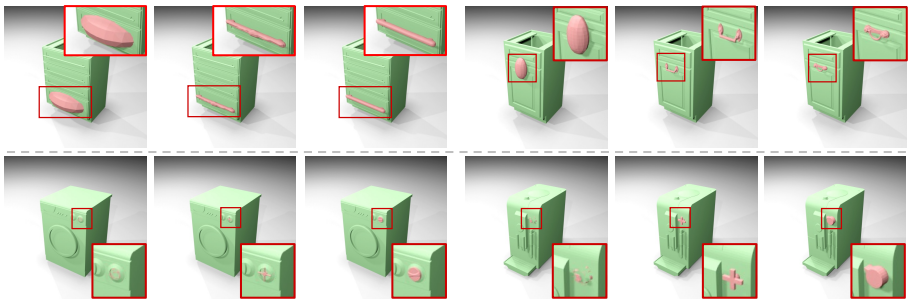

Fig. 5: Results of dependent part generation. All visualizations are done in the objects' canonical poses, with the base shapes in green and the generated parts in red.

Table 2: Quantitative generation results on dependent parts. "Phy." denotes our physical plausibility metric.

Method	Line handle			Hinge knob		
	CD ↓	F-score ↑	Phy. (%)	CD ↓	F-score ↑	Phy. (%)
3D-CVAE	0.01000	0.403	63.3	0.00028	0.894	31.5
Ours (w/o loss guidance)	0.00174	0.742	71.1	0.00026	0.912	78.9
Ours (w/ loss guidance)	**0.00173**	**0.744**	**85.9**	**0.00024**	**0.917**	**84.2**

4.2 Downstream Applications

3D Printing. To better illustrate the physical plausibility of our generated parts, we 3D-printed the generated meshes. The printed parts not only fit seamlessly into the corresponding articulated objects but also demonstrated highly effective motions. Figure 6(a) shows these 3D-printed shapes, which were scaled down to $\frac{1}{200}$ of their original size, resulting in dimensions of approximately 10cm $\times$ 10cm $\times$ 10cm.

Sequential Part Generation. We then explore our method's potential in sequential part generation. For an object with multiple missing parts, we first generate the self-moving parts using the ground truth bounding box of each missing part as the condition simultaneously. Once all self-moving parts are generated, they are assembled with the original object. A point cloud is then re-sampled from this assembly and serves as the condition to generate the dependent parts.

Fig. 6(b) shows our results. Although only trained on objects with one missing part, our model presents outstanding generalizability to objects with multiple missing parts. This capability allows for the sequential part generation that adheres to physical constraints, showcasing real-world applications such as obtaining a hierarchy of physically plausible parts given only bounding box-level design.

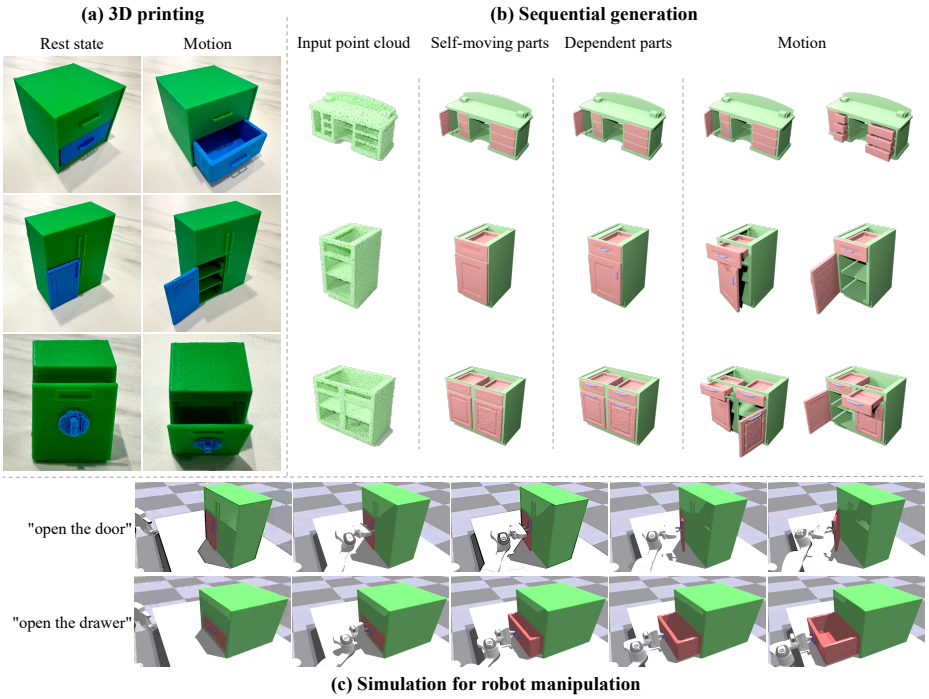

Fig. 6: Downstream applications of our method: (a) 3D printing for real-world object completion, (b) sequential part generation for complex structures, and (c) simulation for robot manipulation.

Simulation for Robot Manipulation. Finally, we demonstrate that the generated parts can be manipulated using simple heuristics to validate the physical plausibility and quality of our generated parts. Following the setup in [16, 17], we load the objects in URDF format into the IsaacGym simulator [41] and apply a heuristic policy based on the estimated pose. Fig. 6(c) shows that our generated shapes exhibit high quality and can be successfully manipulated.

4.3 Ablation Studies and Analysis

This section delves into the effectiveness of different model components through ablation studies. Please refer to Appendix B.1 for additional results and analysis,
Bounding Box Condition. We first verify the design of adding the bounding box condition to the diffusion model. As shown by the results in Tabs. 3 and 4, and Fig. 9, relying solely on the point cloud condition often yields part generation results that fail to fit the input object, such as drawers colliding or doors not filling the whole space. Conversely, incorporating the bounding box as an auxiliary condition significantly improves the overall performance of our model.

Table 3: Ablation studies on the pose proposal model and the conditional generation model on *self-moving* parts.

Method	Slider drawer			Hinge door		
	CD ↓	F-score ↑	Phy. (%)	CD ↓	F-score ↑	Phy. (%)
Regress. pose proposal	**0.00059**	0.847	73.6	0.00134	0.736	50.0
w/o bbox condition	0.00178	0.740	53.5	0.00433	0.791	74.4
Full model	0.00060	**0.862**	**74.3**	**0.00081**	**0.883**	**92.2**

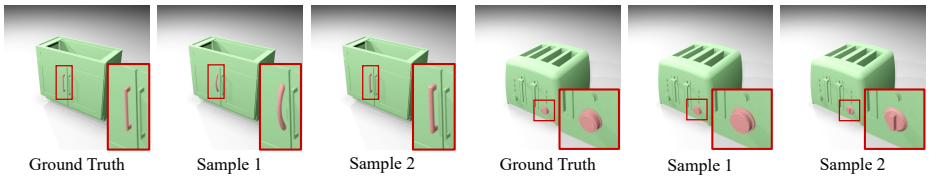

Fig. 7: Illustration of part generation diversity. Despite the strict requirements imposed by the input conditions, our method could generate multiple plausible instances for *dependent* parts.

Pose Proposal Model. We then verify the effectiveness of adopting the diffusion-based pose proposal model. As shown in Tab. 3, leveraging the diffusion-based structure produces object bounding boxes of higher accuracy, thus benefiting the overall part generation quality.

Discussion on Generation Diversity. Using both the object point cloud and predicted part bounding box inherently restricts the diversity of parts that can be generated. This phenomenon is particularly notable for *self-moving* part categories, which must precisely conform to the input object's specifications in size and shape. However, for *dependent* parts like handles and knobs, our model demonstrates the capacity to generate multiple physically plausible part instances, as illustrated in Fig. 7.

5 Conclusions

In this paper, we address the critical challenge of generating physically plausible parts for interactable objects. We propose a diffusion-based framework that integrates geometric conditioning and physical constraints, ensuring high accuracy and fidelity of the generated parts in shape and function under real-world interactions. Our results demonstrate substantial improvements in shape quality and physical plausibility over existing methods. These enhancements significantly benefit downstream applications such as 3D printing and robot manipulation, where precision is crucial for the tasks' success.

Limitations and Future Work. A major limitation is that we only studied articulated objects from common datasets with 1D revolute or prismatic joints. Further studies can be done on more complex joints such as screws and wrenches with joint motions represented by Plücker coordinates. Such complex joint types

can also be coupled with complex part geometries, which are important to certain domains such as tools or mechanical structures. We believe future studies in this direction will facilitate a variety of downstream applications such as CAD design and mechanical engineering.

References

1. Achlioptas, P., Diamanti, O., Mitliagkas, I., Guibas, L.: Learning representations and generative models for 3d point clouds. In: International Conference on Machine Learning (ICML) (2018)
2. Chen, Q., Memmel, M., Fang, A., Walsman, A., Fox, D., Gupta, A.: Urdformer: Constructing interactive realistic scenes from real images via simulation and generative modeling. In: Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition @ CoRL 2023 (2023)
3. Chen, X., Chen, B., Mitra, N.J.: Unpaired point cloud completion on real scans using adversarial training. In: International Conference on Learning Representations (ICLR) (2020)
4. Cheng, K., Wu, R., Shen, Y., Ning, C., Zhan, G., Dong, H.: Learning environment-aware affordance for 3d articulated object manipulation under occlusions. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
5. Cheng, Y.C., Lee, H.Y., Tulyakov, S., Schwing, A.G., Gui, L.Y.: Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
6. Chibane, J., Alldieck, T., Pons-Moll, G.: Implicit functions in feature space for 3d shape reconstruction and completion. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
7. Chung, H., Kim, J., Mccann, M.T., Klasky, M.L., Ye, J.C.: Diffusion posterior sampling for general noisy inverse problems. In: International Conference on Learning Representations (ICLR) (2023)
8. Dai, A., Diller, C., Nießner, M.: Sg-nn: Sparse generative neural networks for self-supervised scene completion of rgb-d scans. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
9. Dai, A., Nießner, M.: Scan2mesh: From unstructured range scans to 3d meshes. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
10. Dai, A., Ruizhongtai Qi, C., Nießner, M.: Shape completion using 3d-encoder-predictor cnns and shape synthesis. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
11. Deng, C., Jiang, C., Qi, C.R., Yan, X., Zhou, Y., Guibas, L., Anguelov, D., et al.: Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
12. Deng, C., Lei, J., Shen, W.B., Daniilidis, K., Guibas, L.J.: Banana: Banach fixed-point network for pointcloud segmentation with inter-part equivariance. In: Advances in Neural Information Processing Systems (NeurIPS) (2024)
13. Ding, Y., Geng, H., Xu, C., Fang, X., Zhang, J., Dai, Q., Wei, S., Zhang, Z., Wang, H.: Open6DOR: Benchmarking open-instruction 6-dof object rearrangement and a VLM-based approach. In: ICRA 2024 Workshop on 3D Visual Representations for Robot Manipulation (2024), https://openreview.net/forum?id=axAmAy3Ghl
14. Firman, M., Mac Aodha, O., Julier, S., Brostow, G.J.: Structured prediction of unobserved voxels from a single depth image. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
15. Geng, H., Li, Z., Geng, Y., Chen, J., Dong, H., Wang, H.: Partmanip: Learning cross-category generalizable part manipulation policy from point cloud observations. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
16. Geng, H., Wei, S., Deng, C., Shen, B., Wang, H., Guibas, L.: Sage: Bridging semantic and actionable parts for generalizable articulated-object manipulation under language instructions. In: Robotics: Science and Systems (RSS) (2024)

17. Geng, H., Xu, H., Zhao, C., Xu, C., Yi, L., Huang, S., Wang, H.: Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
18. Geng, Y., An, B., Geng, H., Chen, Y., Yang, Y., Dong, H.: End-to-end affordance learning for robotic manipulation. International Conference on Robotics and Automation (ICRA) (2023)
19. Gong, R., Huang, J., Zhao, Y., Geng, H., Gao, X., Wu, Q., Ai, W., Zhou, Z., Terzopoulos, D., Zhu, S.C., et al.: Arnold: A benchmark for language-grounded task learning with continuous states in realistic 3d scenes. In: International Conference on Computer Vision (ICCV) (2023)
20. Han, X., Li, Z., Huang, H., Kalogerakis, E., Yu, Y.: High-resolution shape completion using deep neural networks for global structure and local geometry inference. In: International Conference on Computer Vision (ICCV) (2017)
21. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
22. Hu, J., Fei, B., Xu, B., Hou, F., Yang, W., Wang, S., Lei, N., Qian, C., He, Y.: Topology-aware latent diffusion for 3d shape generation. arXiv preprint arXiv:2401.17603 (2024)
23. Huang, Q.: Enhancing implicit shape generators using topological regularizations. In: Joint Mathematics Meetings (JMM 2024) (2024)
24. Huang, S., Wang, Z., Li, P., Jia, B., Liu, T., Zhu, Y., Liang, W., Zhu, S.C.: Diffusion-based generation, optimization, and planning in 3d scenes. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
25. Jiang, Z., Hsu, C.C., Zhu, Y.: Ditto: Building digital twins of articulated objects from interaction. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
26. Kazhdan, M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction. In: Proceedings of Eurographics Symposium on Geometry Processing (2006)
27. Kazhdan, M., Hoppe, H.: Screened poisson surface reconstruction. ACM Transactions on Graphics (TOG) **32**(3), 1–13 (2013)
28. Kim, Y.M., Mitra, N.J., Yan, D.M., Guibas, L.: Acquiring 3d indoor environments with variability and repetition. ACM Transactions on Graphics (TOG) **31**(6), 1–11 (2012)
29. Lei, J., Deng, C., Shen, B., Guibas, L., Daniilidis, K.: Nap: Neural 3d articulation prior. arXiv preprint arXiv:2305.16315 (2023)
30. Li, P., Liu, T., Li, Y., Han, M., Geng, H., Wang, S., Zhu, Y., Zhu, S.C., Huang, S.: Ag2manip: Learning novel manipulation skills with agent-agnostic visual and action representations (2024), `https://arxiv.org/abs/2404.17521`
31. Li, X., Wang, H., Yi, L., Guibas, L.J., Abbott, A.L., Song, S.: Category-level articulated object pose estimation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
32. Li, X., Zhang, M., Geng, Y., Geng, H., Long, Y., Shen, Y., Zhang, R., Liu, J., Dong, H.: Manipllm: Embodied multimodal large language model for object-centric robotic manipulation (2023)
33. Li, Y., Dai, A., Guibas, L., Nießner, M.: Database-assisted object retrieval for real-time 3d reconstruction. In: Computer Graphics Forum (2015)
34. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2023)

35. Liu, G., Sun, Q., Huang, H., Ma, C., Guo, Y., Yi, L., Huang, H., Hu, R.: Semi-weakly supervised object kinematic motion prediction. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
36. Liu, J., Savva, M., Mahdavi-Amiri, A.: Survey on modeling of articulated objects. arXiv preprint arXiv:2403.14937 (2024)
37. Liu, J., Tam, H.I.I., Mahdavi-Amiri, A., Savva, M.: Cage: Controllable articulation generation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
38. Liu, M., Xu, C., Jin, H., Chen, L., Varma T, M., Xu, Z., Su, H.: One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
39. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: International Conference on Computer Vision (ICCV) (2023)
40. Luo, R., Yu, H.X., Wu, J.: Unsupervised discovery of object-centric neural fields. arXiv preprint arXiv:2402.07376 (2024)
41. Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., Hoeller, D., Rudin, N., Allshire, A., Handa, A., State, G.: Isaac gym: High performance gpu-based physics simulation for robot learning (2021)
42. Mezghanni, M., Bodrito, T., Boulkenafed, M., Ovsjanikov, M.: Physical simulation layer for accurate 3d modeling. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
43. Mezghanni, M., Boulkenafed, M., Lieutier, A., Ovsjanikov, M.: Physically-aware generative network for 3d shape modeling. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
44. Mitra, N.J., Guibas, L.J., Pauly, M.: Partial and approximate symmetry detection for 3d geometry. ACM Transactions on Graphics (TOG) **25**(3), 560–568 (2006)
45. Mittal, P., Cheng, Y.C., Singh, M., Tulsiani, S.: Autosdf: Shape priors for 3d completion, reconstruction and generation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
46. Mu, J., Qiu, W., Kortylewski, A., Yuille, A., Vasconcelos, N., Wang, X.: A-sdf: Learning disentangled signed distance functions for articulated shape representation. In: International Conference on Computer Vision (ICCV) (2021)
47. Nan, L., Xie, K., Sharf, A.: A search-classify approach for cluttered indoor scene understanding. ACM Transactions on Graphics (TOG) **31**(6), 1–10 (2012)
48. Nealen, A., Igarashi, T., Sorkine, O., Alexa, M.: Laplacian mesh optimization. In: Proceedings of International Conference on Computer Graphics and Interactive Techniques in Australasia and Southeast Asia (2006)
49. Nguyen, D.T., Hua, B.S., Tran, K., Pham, Q.H., Yeung, S.K.: A field model for repairing 3d shapes. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
50. Pauly, M., Mitra, N.J., Wallner, J., Pottmann, H., Guibas, L.J.: Discovering structural regularity in 3d geometry. ACM Transactions on Graphics (TOG) **27**(3), 1–11 (2008)
51. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
52. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)

53. Rao, Y., Nie, Y., Dai, A.: Patchcomplete: Learning multi-resolution patch priors for 3d shape completion on unseen categories. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
54. Shi, R., Chen, H., Zhang, Z., Liu, M., Xu, C., Wei, X., Chen, L., Zeng, C., Su, H.: Zero123++: a single image to consistent multi-view diffusion base model. arXiv preprint arXiv:2310.15110 (2023)
55. Sipiran, I., Gregor, R., Schreck, T.: Approximate symmetry detection in partial 3d meshes. In: Computer Graphics Forum (2014)
56. Smith, E.J., Meger, D.: Improved adversarial systems for 3d object generation and reconstruction. In: Conference on Robot Learning (CoRL) (2017)
57. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: Advances in Neural Information Processing Systems (NeurIPS) (2015)
58. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (ICLR) (2021)
59. Song, J., Zhang, Q., Yin, H., Mardani, M., Liu, M.Y., Kautz, J., Chen, Y., Vahdat, A.: Loss-guided diffusion models for plug-and-play controllable generation. In: International Conference on Machine Learning (ICML) (2023)
60. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic scene completion from a single depth image. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
61. Sorkine, O., Cohen-Or, D.: Least-squares meshes. In: Proceedings Shape Modeling Applications (2004)
62. Speciale, P., Oswald, M.R., Cohen, A., Pollefeys, M.: A symmetry prior for convex variational 3d reconstruction. In: European Conference on Computer Vision (ECCV) (2016)
63. Sundaralingam, B., Hari, S.K.S., Fishman, A., Garrett, C., Wyk, K.V., Blukis, V., Millane, A., Oleynikova, H., Handa, A., Ramos, F., Ratliff, N., Fox, D.: curobo: Parallelized collision-free minimum-jerk robot motion generation (2023)
64. Sung, M., Kim, V.G., Angst, R., Guibas, L.: Data-driven structural priors for shape completion. ACM Transactions on Graphics (TOG) **34**(6), 1–11 (2015)
65. Tatarchenko, M., Richter, S.R., Ranftl, R., Li, Z., Koltun, V., Brox, T.: What do single-view 3d reconstruction networks learn? In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
66. Thrun, S., Wegbreit, B.: Shape from symmetry. In: International Conference on Computer Vision (ICCV) (2005)
67. Tseng, W.C., Liao, H.J., Yen-Chen, L., Sun, M.: Cla-nerf: Category-level articulated neural radiance field. In: International Conference on Robotics and Automation (ICRA) (2022)
68. Wan, W., Geng, H., Liu, Y., Shan, Z., Yang, Y., Yi, L., Wang, H.: Unidexgrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. arXiv preprint arXiv:2304.00464 (2023)
69. Wang, J., Xu, H., Xu, J., Liu, S., Wang, X.: Synthesizing long-term 3d human motion and interaction in 3d scenes. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
70. Wu, R., Chen, X., Zhuang, Y., Chen, B.: Multimodal shape completion via conditional generative adversarial networks. In: European Conference on Computer Vision (ECCV) (2020)
71. Xu, D., Jiang, Y., Wang, P., Fan, Z., Wang, Y., Wang, Z.: Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360deg views. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2023)

72. Xu, Z., He, Z., Song, S.: Universal manipulation policy network for articulated objects. IEEE Robotics and Automation Letters (RA-L) **7**(2), 2447–2454 (2022)
73. Yang, Y., Jia, B., Zhi, P., Huang, S.: Physcene: Physically interactable 3d scene synthesis for embodied ai. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
74. Yi, L., Huang, H., Liu, D., Kalogerakis, E., Su, H., Guibas, L.: Deep part induction from articulated object pairs. arXiv preprint arXiv:1809.07417 (2018)
75. Yu, X., Zhang, J., Liu, H., Li, D., Geng, H., Wang, H., Ding, Y., Chen, J.: Dexgraspnet 2.0: Learning generative dexterous grasping in large-scale synthetic cluttered scenes. In: 2nd Workshop on Dexterous Manipulation: Design, Perception and Control (RSS) (2024), `https://openreview.net/forum?id=V6XHKFicwc`
76. Yu, X., Rao, Y., Wang, Z., Liu, Z., Lu, J., Zhou, J.: Pointr: Diverse point cloud completion with geometry-aware transformers. In: International Conference on Computer Vision (ICCV) (2021)
77. Yuan, Y., Song, J., Iqbal, U., Vahdat, A., Kautz, J.: Physdiff: Physics-guided human motion diffusion model. In: International Conference on Computer Vision (ICCV) (2023)
78. Zhang, J., Wu, M., Dong, H.: Genpose: Generative category-level object pose estimation via diffusion models. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
79. Zhang, J., Chen, X., Cai, Z., Pan, L., Zhao, H., Yi, S., Yeo, C.K., Dai, B., Loy, C.C.: Unsupervised 3d shape completion through gan inversion. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
80. Zhang, S., Zhang, Y., Ma, Q., Black, M.J., Tang, S.: Place: Proximity learning of articulation and contact in 3d environments. In: International Conference on 3D Vision (3DV) (2020)
81. Zhang, Y., Hassan, M., Neumann, H., Black, M.J., Tang, S.: Generating 3d people in scenes without people. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
82. Zhao, W., Gao, S., Lin, H.: A robust hole-filling algorithm for triangular mesh. The Visual Computer **23**, 987–997 (2007)
83. Zheng, X., Liu, Y., Wang, P., Tong, X.: Sdf-stylegan: Implicit sdf-based stylegan for 3d shape generation. In: Computer Graphics Forum (2022)
84. Zhou, Z., Tulsiani, S.: Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2023)

Algorithm 1: Physical Plausibility Metric

input : Object's point cloud $\mathcal{P}$, predicted missing part's mesh x, set of position candidate C_1 determined by grid search, set of small movement C_2, and loss margin β.
output : Whether the predicted mesh x is physically plausible
initiate : $C_0 = \emptyset$

Remove components in x with volume $< 10^{-7}$;
if *x contains multiple components or is not watertight* **then**
  return false;
Transform the mesh x into SDF representation X;
forall *position candidate* $c \in C_1$ **do**
  $P^c \leftarrow P - c$; `// translate point cloud by` $-c$
  if $\ell_{collision-m}(X, P^c) < \beta$ **and** $\ell_{contact-m}(X, P^c) < \beta$ **then**
    Add c to C_0;
forall $c' \in C_2$ **do**
  if exists $c \in C_0$ **where** $(\ell_{collision-m}(X, P^{c+c'}) \geq \beta$ **or** $\ell_{contact-m}(X, P^{c+c'}) \geq \beta)$ **then**
    return true;
return false;

A Implementation Details

A.1 Physical-Aware Losses and Metric

Physical-Aware Losses. We scale all parts to the unit cube with padding 0.2 before transforming into SDF. The threshold α for both physical-aware losses is 0.005.

Physical-Plausible Metric. We detail the methodology of our proposed physical plausibility metric as follows. We first preprocess the meshes and discard those not watertight or have multiple large connected parts deemed physically infeasible. For the viable meshes, we employ a grid search with a grid length of 0.005 around the part's ground truth position to identify all positions where both collision loss (Eq. (5)) and contact loss (Eq. (6)) fall below a predefined threshold of $\beta = 0.04$. A part is then considered physically plausible if moving it along the $-z$ axis for a distance of 0.02 (applicable to hinge doors) or along the $+x$ and $-x$ axes for a distance of 0.04 (other parts categories) results in the losses exceeding the threshold. Note that the hyperparameters are chosen to ensure that more than 95% of each category's ground truth meshes are physically feasible. The pseudo-code for this algorithm is provided in Algorithm 1.

A.2 Experimental Configuration

Pose Proposal Module. The Pose Proposal Module is trained separately for each part category. We train for 100 epochs on a single NVIDIA RTX 4090 with batch size 8. We adopt the Adam optimizer and a 10^{-4} learning rate. The architecture is the same as GenPose [78].
3D Vector Quantizer. The 3D-VQVAE model is trained using all part instances from the GAPartNet [17] dataset and is shared across all part categories. We train the model for 150000 iterations on 8 NVIDIA V100 GPUs with batch size 1. We adopt the Adam optimizer and cosine learning rate scheduler with a base learning rate 10^{-5}. Detailed architecture is shown in Fig. 8.
Part Generation Module. The part generation diffusion module is trained separately for each part category. We train the model for 100000 iterations on 8 NVIDIA V100 GPUs with batch size 4. We adopt the Adam optimizer and cosine learning rate scheduler with a base learning rate 10^{-5}. Guidance weights are set to $\omega_1 = 3$ and $\omega_2 = 1$ during loss-guided sampling. Fig. 8 illustrates the details.
Part Manipulation Interaction Policy. (1) Slider Drawer: To retrieve items from an open drawer, the gripper moves along the z-axis. To open a drawer, it approaches along the x-axis, typically aiming to grasp a handle on the drawer's front face. (2) Hinge Door: For hinge doors with front-facing handles, the gripper grabs the handle to open the door. After grasping the handle, it rotates around the predicted hinge axis to complete the opening or closing action. For doors without handles, if the door is ajar, the gripper clamps the outer edge along the y-axis of the bounding box to open it. (3) Hinge Knob: For hinge knobs, the gripper clamps the knob like a round handle and rotates the end-effector to perform the task. (4) Fixed Handle: The interaction policy for linear fixed handles mirrors that of round fixed handles. The gripper's opening direction should be perpendicular to the linear fixed handle, aligning parallel to the y-axis of the predicted bounding box.
Part Manipulation Interaction Experiments. We employ a heuristic-based interaction policy to open drawers and doors, manipulate handles, and rotate knobs. Once we determine the part pose, our policy immediately provides the corresponding grasping pose. We then use cuRobo [63] to position our gripper at the grasping pose. Following this, our interaction policy and predicted axis guide the design of the end-effector trajectory to match the part's movement trajectory, interpolating it with a time step of $\frac{1}{200}$. Using Inverse Kinematics and a PID controller, we compute joint poses and maneuver the end-effector along the defined trajectory. Our implementation is independent of ROS, ensuring compatibility with various simulators.

B Additional Experiments

B.1 Additional Quantitative Results

Ablation studies on generating dependent parts. Tab. 4 presents ablation studies on dependent parts generation to further validate our design choice

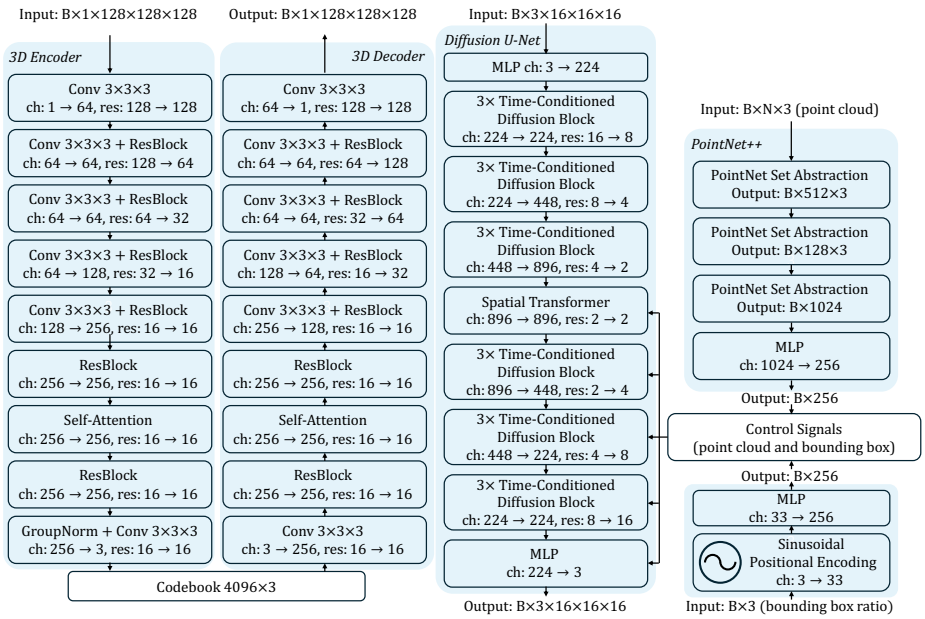

Fig. 8: Architecture of our proposed physically plausible generation framework.

to incorporate bounding box conditions into the diffusion model and employ a diffusion-based architecture for pose prediction. Similar to the results for generating self-moving parts shown in Tab. 3, the inclusion of both designs consistently yields superior results, further justifying the effectiveness of our approach in physical-plausible generation.

B.2 Additional Qualitative Results

Single Part Generation. We present additional single part generation results in Fig. 10 and 11. While the baseline method frequently struggles to generate parts that meet physical constraints, our approach consistently yields high-fidelity part-generation results.

Table 4: Ablation studies on the pose proposal module and the conditional generation module for *dependent* parts.

Method	Line handle			Hinge knob		
	CD ↓	F-score ↑	Phys. (%)	CD ↓	F-score ↑	Phys. (%)
Regress. pose proposal	0.00252	0.655	74.2	0.00036	0.847	81.1
w/o bbox condition	0.00784	0.572	67.0	0.00033	0.885	80.7
Full version	**0.00173**	**0.744**	**85.9**	**0.00024**	**0.917**	**84.2**

Fig. 9: Comparison between whether or not using the bounding box as an additional condition. Incorporating the bounding box beside the point cloud enhances the fit of the generated part to the given object.

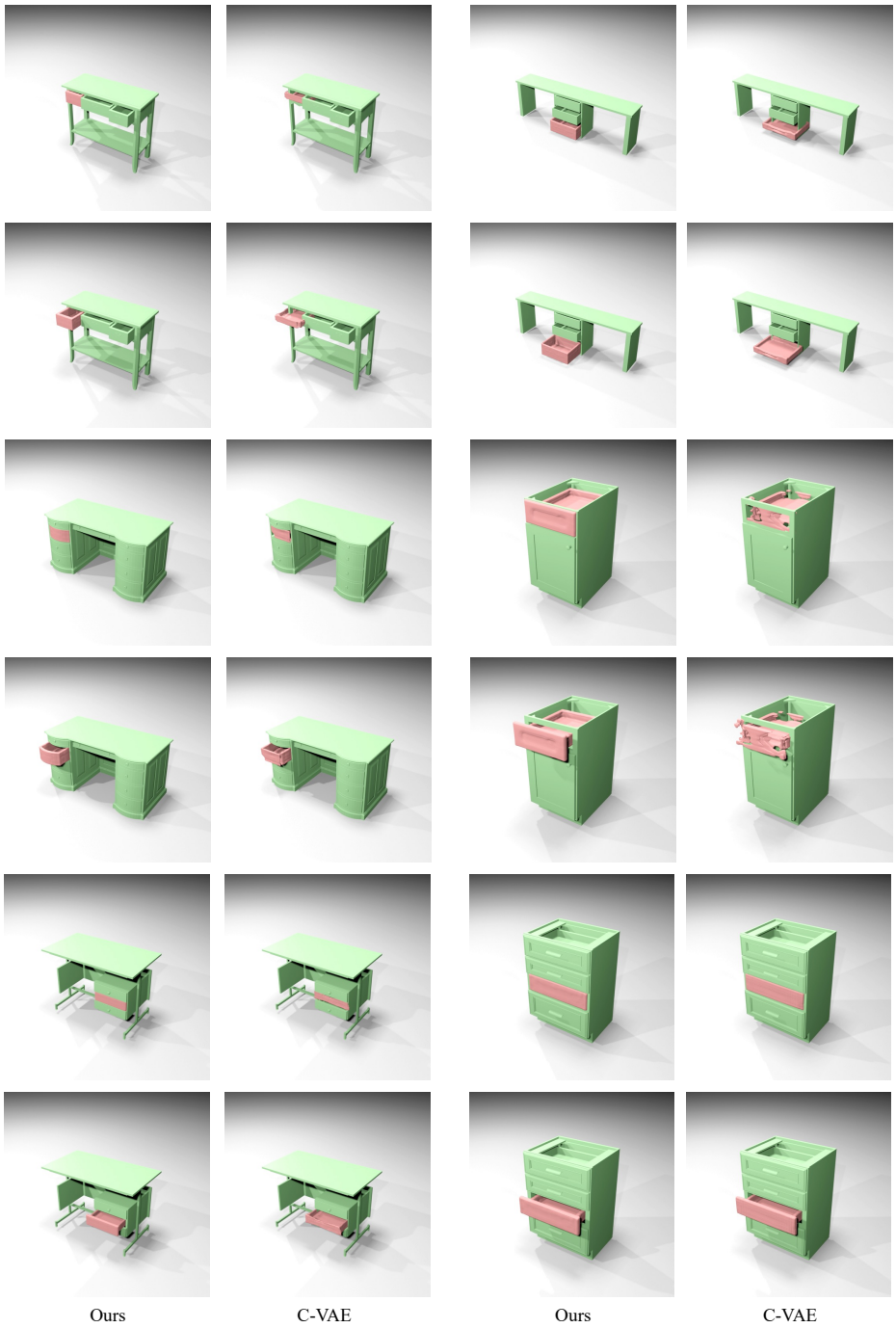

Fig. 10: Qualitative single part generation results.

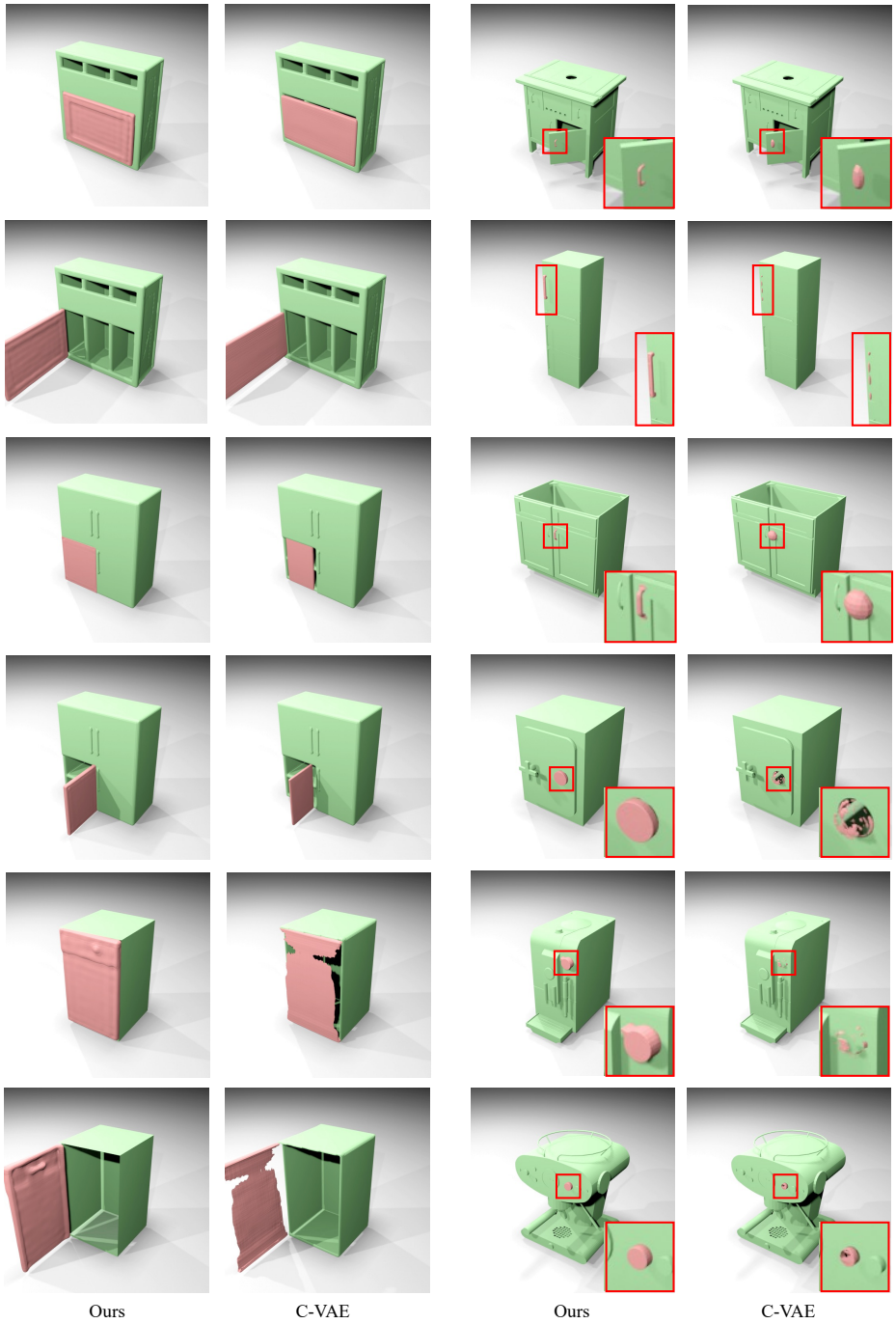

Fig. 11: Qualitative single part generation results.