TOSC: Task-Oriented Shape Completion for Open-World Dexterous Grasp Generation from Partial Point Clouds

Weishang Wu[1], **Yifei Shi**[1*], **Zhiping Cai**[1*]

[1]National University of Defense Technology
wuweishang24@nudt.edu.cn, yifei.j.shi@gmail.com, zpcai@nudt.edu.cn

Abstract

Task-oriented dexterous grasping remains challenging in robotic manipulations of open-world objects under severe partial observation, where significant missing data invalidates generic shape completion. In this paper, to overcome this limitation, we study *Task-Oriented Shape Completion*, a new task that focuses on completing the potential contact regions rather than the entire shape. We argue that shape completion for grasping should be explicitly guided by the downstream manipulation task. To achieve this, we first generate multiple task-oriented shape completion candidates by leveraging the zero-shot capabilities of object functional understanding from several pre-trained foundation models. A 3D discriminative autoencoder is then proposed to evaluate the plausibility of each generated candidate and optimize the most plausible one from a global perspective. A conditional flow-matching model named FlowGrasp is developed to generate task-oriented dexterous grasps from the optimized shape. Our method achieves state-of-the-art performance in task-oriented dexterous grasping and task-oriented shape completion, improving the Grasp Displacement and the Chamfer Distance over the state-of-the-art by 16.17% and 55.26%, respectively. In particular, it shows good capabilities in grasping objects with severe missing data. It also demonstrates good generality in handling open-set categories and tasks.

Code — https://github.com/SyKszzzzz/TOSC

Introduction

Generating dexterous grasps that are both stable and effective for specific downstream tasks remains a core challenge in robot manipulation, attracting substantial research attention. Task-oriented dexterous grasping addresses this challenge by generating grasp poses explicitly designed to facilitate the intended manipulation (Zhong et al. 2025; Wei et al. 2024). This capability is essential for advancing versatile robotic applications in fields such as household service and industrial automation (She et al. 2024; Liu et al. 2025).

Recent advances in large foundation models have enabled zero-shot task-oriented dexterous grasping for open-world objects. While existing methods show promise on synthetic objects or real objects with complete geometry (Li et al. 2024b; Zhong et al. 2025; Li et al. 2024a; Mirjalili et al. 2024; Jian et al. 2025), their performance significantly deteriorates in cluttered real-world environments, due to the influences of severe occlusion, background clutter, and sensor noise. This limitation essentially stems from the requirements of detailed functional understanding, which are unattainable without knowing the completed geometry.

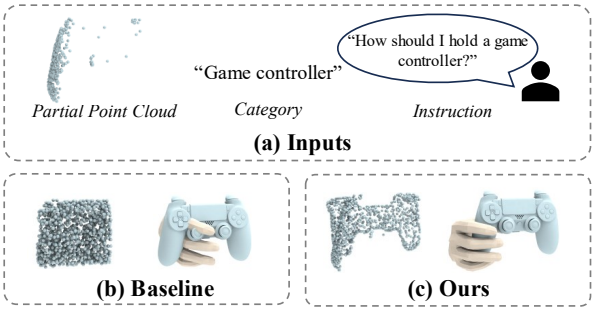

Figure 1: By targeting task-oriented shape completion instead of generic shape completion, our method achieves higher completion accuracy, enabling more plausible task-oriented grasps, compared to the baseline (Wei et al. 2024).

A straightforward solution to address this problem is first to perform a shape completion on the input data, then generate grasps based on the completed shape (Iwase et al. 2025; Kim et al. 2025). However, this decoupled approach would yield incorrect estimation of shape geometry for both the entire object and its contact regions, primarily due to the ambiguity caused by data missing. Consequently, grasps generated on these erroneous completions might fail to satisfy the requirements of the specific grasp tasks.

To solve this problem, this paper studies *Task-Oriented Shape Completion*, a new task focusing exclusively on reconstructing contact regions rather than full object geometry. Our core insight is that shape completion for grasping should be explicitly conditioned on the downstream manipulation task (see Figure 1). As a consequence, the completion process is task-aware and capable of generating completed shapes that facilitate the execution of the specific manipulation task, tolerating the imperfections in irrelevant regions.

To this end, we propose a method that consists of several

*Corresponding author: Yifei Shi, Zhiping Cai

key components. First, it generates multiple candidates of task-oriented completed shapes by leveraging the zero-shot capabilities of object functional understanding from pre-trained foundation models. Specifically, it first synthesizes multiple plausible RGB images containing the potential contact regions conditioned on the input point cloud through the ControlNet (Zhang, Rao, and Agrawala 2023). It then generates 3D shapes from the RGB images by adopting a 3D shape generation network (Zhao et al. 2025).

Second, the generated shapes may be imperfect due to hallucinations during RGB image synthesis and incorrect estimations during shape generation. To alleviate this problem, we propose a 3D discriminative autoencoder (DAE) to select the optimal generated shape and further optimize its geometry from a global perspective. The 3D DAE is trained on large-scale datasets with a well-designed training data generation scheme, enabling it to accurately restore the shape of open-set categories.

Third, we proposed FlowGrasp, a conditional flow-matching model to generate task-oriented dexterous grasps from the optimal shape. Specifically, in the standard flow-matching framework, we apply a single-step, input-side gradient correction to each predicted velocity to implicitly enforce geometric and semantic constraints. Then merge the corrected velocity with the original flow target to update the model. This requires no additional explicit losses or inference overhead, yet guides the model toward constraint-aware optimization.

We conduct extensive experiments to evaluate the effectiveness of the proposed method. Our method achieves state-of-the-art performance in task-oriented dexterous grasping and task-oriented shape completion, improving the Grasp Displacement and the Chamfer Distance over the state-of-the-art by 16.17% and 55.26%, respectively. In particular, it shows good capabilities in grasping objects with severe missing data. It also demonstrates good generality in handling open-set categories and tasks.

In summary, the contributions of this paper are as follows:

- We propose task-oriented shape completion, a new task that focuses exclusively on completing contact regions rather than the full object geometry.
- We develop a method that first generates multiple task-oriented shape completion candidates by several pre-trained foundation models and then selects the most plausible one as well as optimizes its geometry from a global perspective by developing a 3D DAE.
- We introduce a constraint-aware conditional flow-matching model that enforces geometric and semantic constraints via a single-step gradient correction, with no extra losses or inference overhead.
- Our method achieves state-of-the-art performance on task-oriented dexterous grasping and shape completion.

Related Work

Point Cloud Completion

Point cloud completion is a long-standing research topic. Recent advances in point cloud completion are dominated by learning-based methods. Most methods directly estimate the missing part in a single forward pass by training on a large number of labeled data (Yuan et al. 2018; Yu et al. 2021). Despite the satisfactory performance, the performance of these methods relies heavily on the quality of training data (Li, Zhu, and Wei 2025) and has limited capability in out-of-distribution data. Another line of work integrates external 2D/3D priors to improve the generality (Kasten, Rahamim, and Chechik 2023; Huang et al. 2024; Li, Zhu, and Wei 2025). Our method is inspired by the existing methods. However, our method explicitly incorporates the manipulation task to guide the completion process, making it a more practicable solution for robot manipulation.

Task-Oriented Grasping

Task-oriented grasping is a crucial yet challenging task that requires generating grasps not only for stability but also to enable the execution of a subsequent task. Existing methods aggregate features from language and vision inputs, providing zero-shot capability to both novel instruction and category (Wei et al. 2024; Li et al. 2024b). For example, DexTOG (Zhang et al. 2024) learns intermediate features for grasp-instruction alignment. DexGraspVLA (Zhong et al. 2025) proposes a hierarchical vision-language-action framework to implement grasp generation. AffordDexGrasp (Wei et al. 2025) proposes a generalizable affordance representation that aligns robot actions with language semantics. However, these methods require complete shapes as input, greatly degrading their effectiveness in real-world scenarios with partial observation. To address this issue, several approaches that first complete the input shape and then generate grasps based on the completed one are proposed (Kim et al. 2025; Iwase et al. 2025). Meanwhile, approaches that rely on massive amounts of supervised data can indeed partly overcome the limitations of single-view inputs (Zhong et al. 2025; Feng et al. 2024), but they incur very high data-collection and annotation costs. Moreover, their ability to generalize to unseen object categories or to succeed under extreme occlusions and atypical geometries has yet to be rigorously demonstrated. Our method improves upon prior work by replacing generic point cloud completion with task-oriented shape completion, substantially enhancing the performance of task-oriented grasp generation from partial point clouds.

Masked Autoencoders for Point Clouds

3D masked autoencoders (MAEs) have recently emerged as a powerful self-supervised paradigm for pre-training on 3D geometric data. By restoring 3D structures from partially masked inputs, these models learn rich representations without explicit supervision. The seminal Point-MAE (Pang et al. 2023) pioneered this approach for point clouds, establishing the mask and reconstruction pre-training scheme. Subsequent works enhance MAEs for point clouds through various strategies (Zhang et al. 2022; Zha et al. 2024; Zhang, Zhang, and Yan 2024). While building upon a similar architecture for shape restoration, our method extends the existing works by introducing an extra discriminative component. This addition enables the identification of implausible input shapes, allowing us to filter incorrect estimations.

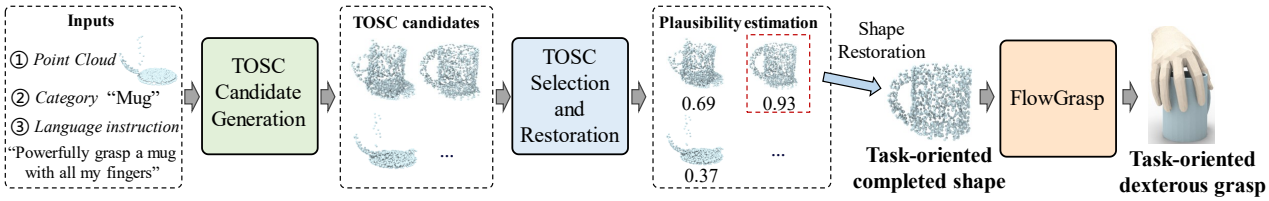

Figure 2: The overview of our method. Taking a partial point cloud of an object, the object's category, and a language description of a manipulation task as input, our method first generates multiple candidates of task-oriented completed shapes by the TOSC candidate generation. It then evaluates the plausibility of the generated candidates and restores the most plausible shape from a global perspective by the TOSC selection and restoration. Last, the task-oriented dexterous grasp is generated by the FlowGrasp.

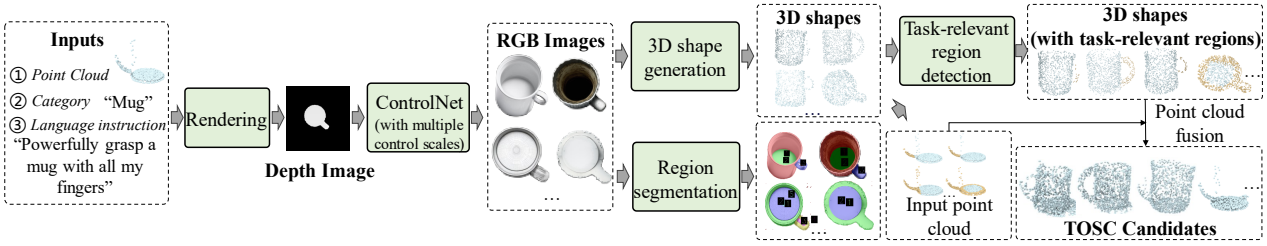

Figure 3: The pipeline of the TOSC candidate generation. First, the input point cloud is rendered into a depth map. Then, the ControlNet is adopted to synthesize multiple RGB images using different control scales. Third, the corresponding 3D shapes are then generated with a 3D shape generation network. After segmenting and detecting task-relevant regions in the generated 3D shapes and input point cloud, a point cloud fusion is performed to generate the TOSC candidates.

Method

Overview

Taking the partial point cloud $P_{\text{in}} \in \mathbb{R}^{N_{\text{in}} \times 3}$ of a 3D object, its category label C, and the language description of a manipulation task G as input, our method completes the shape and generates dexterous grasps as follows. First, it generates multiple candidates of task-oriented completed shapes by leveraging several pre-trained foundation models. Second, it evaluates the plausibility of the generated candidates and optimizes the most plausible shape. Third, the task-oriented grasp is generated by a conditional Flow-Matching grasp generation network. An overview is visualized in Figure 2.

TOSC Candidate Generation

Despite the recent progress in shape completion, most existing methods complete the geometry based solely on input point clouds, neglecting their relevance to downstream tasks. These methods would yield incorrect estimations, due to the ambiguity caused by data missing. Consequently, grasps generated on these erroneous completions might fail to satisfy the requirements of the specific manipulation tasks.

To alleviate this problem, we opt to task-oriented shape completion, which explicitly conditions shape completion on downstream manipulation tasks. As such, the completion is task-aware, i.e., it focuses on completion of regions that could facilitate the execution of the specific manipulation task, tolerating the imperfections in irrelevant regions.

Direct learning to estimate potential contact regions purely from input point clouds is non-trivial due to the lack of open-world knowledge regarding manipulation tasks. To solve this problem, we introduce a method that first generates multiple plausible RGB images from the input point cloud by leveraging a pre-trained vision language model. The generated RGB images correspond to multiple plausible shapes, allowing our method to handle ambiguity caused by incomplete observation. The 3D objects corresponding to each RGB image are then generated by using a pre-trained 3D generative model.

Specifically, as shown in Figure 3, the input point cloud $P_{\text{in}} \in \mathbb{R}^{N_{\text{in}} \times 3}$ is rendered into a depth image I_{depth}. To select the viewpoint for rendering, we adopt the Hidden-Point-Removal (Katz, Tal, and Basri 2007), which reformulates viewpoint estimation as a hidden point removal task to generate a viewpoint V that maximizes visible points. The depth image is then fed into the ControlNet (Zhang, Rao, and Agrawala 2023) as a condition to synthesize RGB images, with the object's category C serving as the prompt. ControlNet injects I_{depth} at multiple U-Net feature levels via zero-initialized convolutional adapters, without requiring any additional task description. To generate multiple plausible RGB images, we employ multiple different control scale λ on the ControlNet branch to control the balance between strict geometric compliance and semantic completion, where a large λ corresponds to "pay more attention" to the input depth map, preserving fine-grained shape details but also faithfully reproducing any missing-region artifacts, and vice versa.

Having generated the RGB images, we then adopt a pre-trained 3D shape generation network (Zhao et al. 2025) G_M to estimate the mesh M for each RGB image, from which a point set $\{P^i_{\text{gen}} \in \mathbb{R}^{N_{\text{gen}} \times 3}\}_{i=0}^{N_{\text{RGB}}}$ is sampled. G_M is a flow-

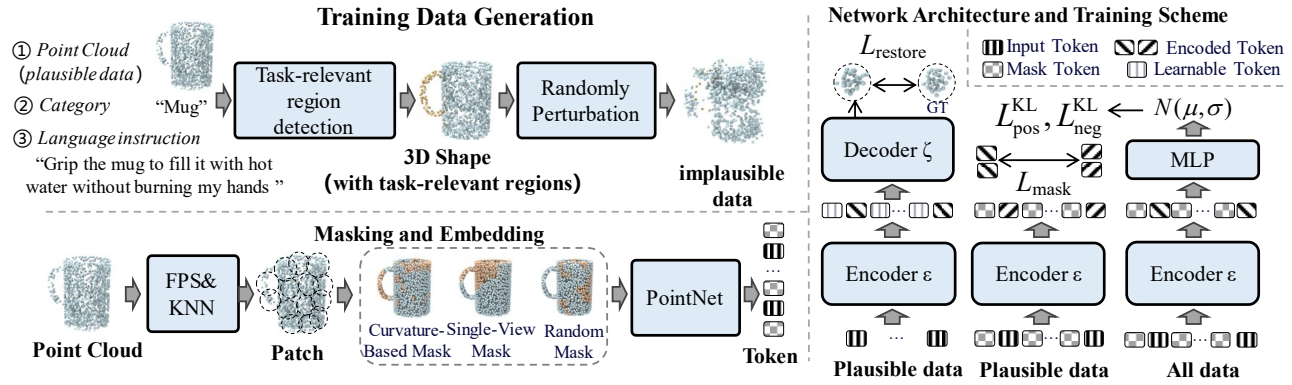

Figure 4: Illustrations of the key components in the TOSC selection and restoration.

based diffusion model, which adopts double- and single-stream blocks, with DINOv2 Gaint encoding RGB image as condition injection.

The generated shapes are not necessarily perfectly aligned with the input point cloud. Hence, we perform an ICP algorithm to align and fuse the two point clouds. Since our method focuses on the completion of task-relevant regions, we first use SAM (Kirillov et al. 2023) to segment each RGB image into multiple regions and then adopt a multi-modal large model (Achiam et al. 2023) to detect the task-relevant regions in the image space. Those regions are then projected onto both the input point cloud and the generated point cloud to acquire $P_{\text{in}}^{\text{task}}$ and $P_{\text{gen}}^{\text{task}}$. We optimizing for the scale k and transformation tr via:

$$\underset{k,tr}{\text{argmin}}[\text{CD}(P_{\text{in}}, tr(kP_{\text{gen}})) + w_{\text{task}}\text{CD}(P_{\text{in}}^{\text{task}}, tr(kP_{\text{gen}}^{\text{task}}))], \tag{1}$$

where $\text{CD}(\cdot,\cdot)$ denotes the chamfer distance, w_{task} making the optimization focuses on aligning the task-relevant regions. We denote the fused point cloud $P_{\text{can}} \in \mathbb{R}^{N_{\text{can}}\times 3}$ as a TOSC candidate.

TOSC Selection and Restoration

The generated 3D shape candidates above may be imperfect due to hallucinations during RGB image synthesis and incorrect estimations of shape reconstruction. To solve this problem, we propose a DAE that jointly evaluates the plausibility of each 3D shape candidate and restores the geometry of the most plausible shape from a global perspective. The diagram is shown in Figure 4.

Training data generation. To effectively train the 3D DAE, we collect a dataset that contains both the plausible and implausible shapes w.r.t the input manipulation task. We define the plausible shapes as those that have sufficient geometry that supports the execution of the input manipulation task. To achieve this, we collect objects from 6 datasets, i.e. ModelNet40 (Wu et al. 2015), ShapeNetCore (Chang et al. 2015), ScanObjectNN (Uy et al. 2019), OmniObject3D (Wu et al. 2023), DexGraspNet (Wang et al. 2022), and AffordPose (Jian et al. 2023). In total, the training set of plausible data contains $72,524$ objects, exhibiting significant variations in both synthetic and real-world scenarios.

To generate implausible data, we deliberately sabotage the plausible data by the following steps. We first pair each object with a randomly selected manipulation task and prompt the large language model to identify the contact region implied by the manipulation task. Then, we apply the PartSlip (Liu et al. 2023) to identify the task-relevant segment from the object. The implausible data are generated by randomly removing the task-relevant segment, adding noise, and perturbing local patches.

Network architecture. In general, the 3D DAE contains an encoder $\varepsilon(\cdot)$ and an decoder $\zeta(\cdot)$: $l_{\text{can}} = \varepsilon(P_{\text{can}}), \hat{P_{\text{can}}} = \zeta(l_{\text{can}})$. P_{can} is the point cloud of 3D shape candidate, l_{can} is the latent vector of P_{can}, $\hat{P_{\text{can}}}$ is the restored point cloud.

The encoder $\varepsilon(\cdot)$ consists of N_{encoder} standard Transformer blocks. To improve the computational efficiency, we tokenize the input P_{can} before feeding it into the encoder ε. Specifically, we first split P_{can} into N_{patch} local patches by a farthest point sampling (FPS) and a K-nearest neighbor (KNN) algorithm. Each patch is then tokenized by a lightweight PointNet (Qi et al. 2017). Inspired by existing works of MAEs (Pang et al. 2023; Wang et al. 2021; Kakogeorgiou et al. 2022), during training, we randomly mask several patch tokens in the input P_{can} to get mask and visible point cloud $P_{\text{can}}^{\text{mask}}$, $P_{\text{can}}^{\text{vis}}$. Note that, for the plausible data, we did not mask the patch tokens in the task-relevant segments.

To evaluate the plausibility of the input P_{can}, we first adopt the latent vector l_{can} to estimate the mean μ and the standard deviation σ of a Gaussian distribution $\mathcal{N}(\mu,\sigma)$ by:

$$\begin{aligned} \mu &= \text{MLP}_{\mu}(l_{\text{can}}), \\ \sigma &= \text{MLP}_{\sigma}(l_{\text{can}}), \end{aligned} \tag{2}$$

where $\text{MLP}_{\mu}(\cdot)$ and $\text{MLP}_{\sigma}(\cdot)$ are fully-connected layers. Then, a Kullback-Leibler divergence loss $L_{\text{pos}}^{\text{KL}}$ is used to optimize the predicted Gaussian distribution $\mathcal{N}(\mu,\sigma)$ of plausible shapes to approximate the standard normal distribution, i.e. $\mathcal{N}(0,1)$, and adopt an $L_{\text{neg}}^{\text{KL}}$ to optimize implausible shapes to approximate another distribution, i.e. $\mathcal{N}(1,1)$.

The decoder $\zeta(\cdot)$ consists of N_{decoder} Transformer blocks. It first aggregates features in the tokens and then generates

the restored point cloud P_{recon} via a fully-connected layer. A chamfer distance loss function between the restored point cloud and the ground-truth is applied:

$$L_{\text{restore}} = \text{CD}(P_{\text{restore}}, P_{\text{GT}}), \tag{3}$$

where P_{GT} represents the ground-truth.

Besides, the features of patch tokens are expected to be similar before and after the patch tokens are masked. To improve the robustness under occlusion, we feed all the patch tokens and the masked patch tokens to the encoder $\varepsilon(\cdot)$ to extract features respectively and penalize their differences:

$$L_{\text{mask}} = \sum_{i=0}^{N_{\text{vis}}} \text{MSE}[\varepsilon_{T_i}(P_{\text{can}}^{\text{vis}}), \varepsilon_{T_i}(P_{\text{can}})], \tag{4}$$

where MSE$[\cdot]$ denotes the mean squared error, $\varepsilon_{T_i}(P_{\text{can}})$ represents the feature of patch token T_i extracted by encoder $\varepsilon(\cdot)$ taking P_{can} as input. Note that the above L_{recon} and L_{mask} are only used for positive data. Overall, the training loss function is: $L = L_{\text{pos}}^{\text{KL}} + L_{\text{neg}}^{\text{KL}} + L_{\text{recon}} + L_{\text{mask}}$.

Network inference. During inference, the encoder processes an input point cloud P_{can} and outputs a latent vector l_{can} as well as a distribution $d_{\text{can}} = \mathcal{N}(\mu, \sigma)$. The shape plausibility can be estimated with the out-of-distribution likelihood estimation. To achieve this, we consider the KL divergence between the prior and posterior distribution of the input point cloud. The shape plausibility s_{can} measured by:

$$s_{\text{can}} = \text{Sigmod}[-\mathcal{D}_{\text{KL}}(d_{\text{can}}||\mathcal{N}(0,1)) + \mathcal{D}_{\text{KL}}(d_{\text{can}}||\mathcal{N}(1,1))], \tag{5}$$

where Sigmod$[\cdot]$ is the sigmoid function, $\mathcal{D}_{\text{KL}}$ is the KL divergence. A high s_{can} represents that P_{can} is closer to the distribution of plausible shapes than that of implausible shapes, indicating P_{can} is plausible. The decoder can then be adopted to generate the restored shape.

FlowGrasp

Having a plausible and restored 3D shape, the next step is to compute the dexterous task-oriented grasp. Conventional methods typically adopt a conditional variational autoencoder or diffusion models and enforce task and physical constraints by adding weighted penalty terms or by performing gradient-based adjustments at inference time (Wei et al. 2024; Jiang et al. 2021), which either undermines the probabilistic model's log-likelihood interpretation or incurs significant computational overhead.

To solve this problem, we introduce FlowGrasp, a constraint-aware conditional flow matching model that integrates both task and physical constraints directly into training via a single-step, input-side gradient correction—without any extra losses or inference-time adjustments, the illustration is show in Figure 5.

Specifically, we form the condition vector l_{con} by concatenating the PointNet++ feature of the restored 3D shape with the CLIP embedding of the task language description. During training, we sample a ground-truth grasp vector x_1 and an initial noise $x_0 \sim \mathcal{N}(0, \text{I})$, pick timestamp $t \in [0, 1]$, and interpolate x_t under a fixed diffusion kernel $x_t = (1-t)x_0 + t \cdot x_1$.

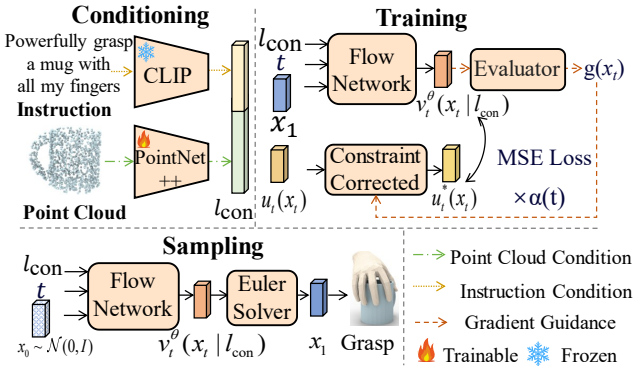

Figure 5: The illustrations of the FlowGrasp.

In the standard flow-matching framework, the instantaneous velocity is $u_t(x_t) = x_1 - x_0$. To enforce constraints, we apply a one-step correction to this velocity:

$$u_t^*(x_t) = u_t(x_t) - \alpha(t)\nabla\left(\sum_i w_{\text{con}}^i g_i(x_t)\right), \tag{6}$$

where each g_i encodes a u or semantic constraint, w_{con} is weighting factor. ∇ is the gradient operator. $\alpha(t)$ is a time-decay factor. We then train the network θ to regress $u_t^*(x_t)$ by minimizing:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{x_0, x_1, t, l_{\text{con}}} ||v_t^\theta(x_t|l_{\text{con}}) - u_t^*(x_t)||^2 \tag{7}$$

During inference, we draw an initial noise vector $x_0 \sim \mathcal{N}(0, \text{I})$ and integrate the learned Ordinary Differential Equation (ODE) $dx/dt = v_t^\theta(x_t|c)$ from $t = 0$ to $t = 1$ to recover the task-oriented dexterous grasp.

Experiments

Implementation Details

The number of points in both the partial and candidate clouds N_{in} and N_{can} is 2048. We convert the depth image to RGB images with ControlNet (Zhang, Rao, and Agrawala 2023) and then reconstruct point clouds from RGB using Hunyuan3D-DiT-v2-mini-Fast (Zhao et al. 2025). Task-relevant region detection is achieved by GPT-4o. FlowGrasp is trained on a single NVIDIA GeForce RTX 4090 GPU for 350 epochs with the Adam optimizer and a batch size of 64. The 3D DAE is trained for 300 epochs with a learning rate of 0.0005 and a weight decay of 0.05.

Experimental Dataset

Our method, except for the 3D DAE, is trained on a new dataset named OakInk-PartialIPC, which containing partial point clouds created from OakInk (Yang et al. 2022). To generate the partial point clouds, we uniformly sample viewpoints around each object and render depth images with random scanning noise and random foreground occlusion to simulate depth maps scanned in real-world environments. The corresponding language instructions are drawn from the CapGrasp dataset (Li et al. 2024b), which provides task descriptions specifically tailored to OakInk.

Figure 6: Visual comparison of the generated task-oriented grasps by ours and the baselines.

Method	Penetration		Grasp Displace		Contact Radio ↑	P-FID ↓	LLM score ↑	Perceptual Score		
	Volume(cm^3) ↓	Depth(cm) ↓	Mean(cm) ↓	Var(cm) ↓				SC↑	PP↑	IS↑
GraspCVAE (Jiang et al. 2021)	16.84	0.141	3.92	4.34	94.74%	39.03	55.0	1.45	1.32	1.27
GraspTTA (Jiang et al. 2021)	16.39	0.159	3.71	4.19	96.25%	38.85	65.0	2.36	1.86	1.99
SceneDiffuser (Huang et al. 2023)	**6.52**	**0.090**	3.81	4.02	95.62%	29.38	61.7	2.27	1.95	1.82
DexTOG (Zhang et al. 2024)	14.32	0.110	3.74	3.78	96.12%	25.93	60.0	2.54	2.22	2.13
DexGYSGrasp (Wei et al. 2024)	7.16	0.096	3.76	3.99	97.20%	25.98	68.3	3.04	2.23	1.99
Ours	6.87	**0.090**	**3.11**	**3.54**	**98.30%**	**21.60**	**88.3**	**4.38**	**3.84**	**3.80**

Table 1: Quantitative comparisons of task-oriented grasping on the OakInk-PartialPC dataset.

Method	CD-$\ell_2 \times 10^{-4}$ ↓	F-Score@1↑	DCD↓
PointAttn (Wang et al. 2024)	4.58	0.512	0.698
SVDFormer (Zhu et al. 2023)	3.71	0.643	0.603
SymmCompletion (Yan et al. 2025)	3.94	0.618	0.611
Ours	**1.66**	**0.860**	**0.488**

Table 2: Quantitative comparisons of point cloud completion on the OakInk-PartialPC dataset.

Evaluation Metrics

To comprehensively evaluate the quality of the generated grasps, we employ multiple evaluation metrics: 1) *Penetration* quantifies the volumetric overlap and maximum insertion depth between the grasp and object (Hasson et al. 2019); 2) *Grasp displacement* evaluates the grasping stability by measuring the displacement of object's mass center under external forces for each grasp; 3) *Contact ratio* reflects the quality of hand–object interaction by calculating sample-level contact ratio (Tzionas et al. 2016); 4) *Point cloud Fréchet Inception Distance (P-FID)* compares high-level feature distributions of generated grasps against ground truth (Nichol et al. 2022); 5) *LLM score* evaluates the quality of grasps with GPT4v (Li et al. 2024b); 6) *Perceptual Scores* are acquired by a user study in which human evaluators each assess 100 distinct grasps and assign scores from 0 (poor) to 5 (excellent) on three criteria: Semantic Consistency (SC), Physical Plausibility (PP), and Interaction Stability (IS).

For evaluation of point cloud completion, we use the standard metrics: 1) Chamfer Distance (CD) (Wang et al. 2024); 2) F1-score (Yan et al. 2025) ; 3) Density-aware CD (DCD) (Wu et al. 2021).

Compare to Baselines

We first compare our method to baselines in task-oriented dexterous grasping. Table 1 reports the quantitative comparisons against several baselines on the OakInk-PartialPC dataset. It shows that our method achieves state-of-the-art performances in metrics of Grasp Displace and competitive performances in metrics of Penetration, demonstrating the superior physical reliability of the generated grasps. Note that the lower penetration volume of SceneDiffuser (Huang et al. 2023) stems from their conservative strategy, which keeps unnecessary distances between hand and object. Our method, in contrast, maintains better balances between penetration avoidance and grasp stability. Moreover, our method achieves the highest LLM Score and perceptual scores, showing its ability to produce high-quality semantically correct grasps that fulfill the language instructions. Qualitative comparisons are visualized in Figure 6.

We then evaluate our method in terms of task-oriented shape completion on the OakInk-PartialPC dataset. Table 2 reports quantitative comparisons with several baselines for generic point cloud completion. The baselines are recent state-of-the-art methods, including PointAttn (Wang et al. 2024), SVDFormer (Zhu et al. 2023), and SymmCompletion (Yan et al. 2025). Our method outperforms all baselines by a large margin in all metrics, revealing the advantages of task-oriented shape completion. Qualitative comparisons are shown in Figure 7.

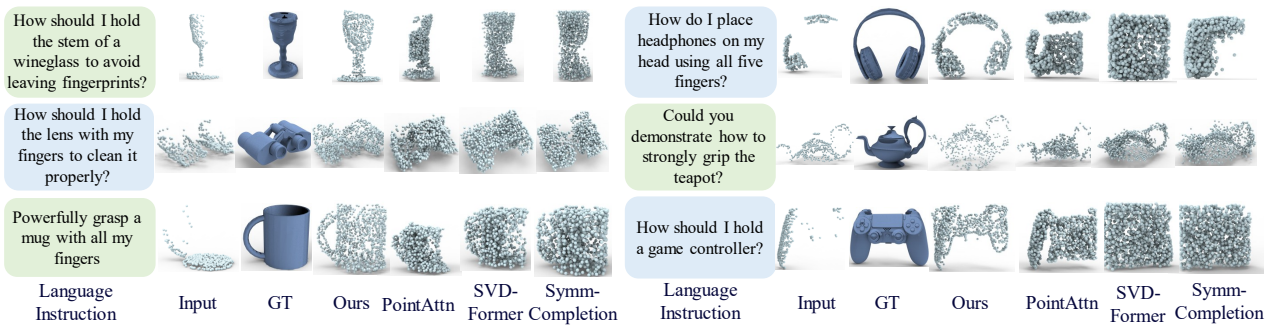

Figure 7: Visual comparisons of point cloud completion.

Method	Penetration		Grasp Displacement		Contact Radio ↑	P-FID ↓	LLM score ↑	Perceptual Score		
	Volume(cm^3) ↓	Depth(cm) ↓	Mean(cm) ↓	Var(cm) ↓				SC↑	PP↑	IS↑
w/o TCG	6.70	0.098	3.50	3.79	97.74%	22.34	71.7	2.63	0.81	1.72
w/o TSR	7.96	0.093	3.35	3.68	98.84%	22.96	75.0	2.36	2.63	2.72
w/o TOSC	7.12	0.090	3.51	3.94	96.79%	23.83	66.7	2.18	1.36	2.18
w/o token masking	7.58	0.087	3.21	3.71	**99.50%**	22.53	78.3	3.09	3.18	3.36
w/o gradient guidance	6.78	**0.086**	3.43	3.70	98.83%	24.01	83.3	3.72	2.18	3.63
The full method	**6.87**	0.090	**3.11**	**3.54**	98.30%	**21.60**	**88.3**	**4.38**	**3.84**	**3.80**

Table 3: Ablation studies of several crucial components in our method.

Method	Penetration		Grasp Displace		P-FID ↓
	Volume(cm^3) ↓	Depth(cm) ↓	Mean(cm) ↓	Var(cm) ↓	
GraspCVAE	16.02	1.18	4.59	4.77	57.01
GraspTTA	14.42	**1.04**	4.59	4.83	53.58
SceneDiffuser	12.43	1.16	4.53	4.70	46.51
DexTOG	18.72	1.31	4.42	4.80	50.53
DexGYSGrasp	12.60	1.28	4.77	4.54	46.42
Ours	**11.53**	1.32	**4.21**	**4.52**	**42.97**

Table 4: Quantitative comparisons of task-oriented grasping on novel category.

Evaluation of Zero-shot Generality

We first evaluate the zero-shot generalization in terms of handling novel object categories and novel language instructions. We experiment on 9 novel object categories. Each category contains about 100 objects and several novel language instructions. As shown in Table 4, our method achieves the best performance across all metrics. It demonstrates that our method can robustly handle novel categories and untrained language instructions in open-world scenarios, thanks to our design of using pre-trained foundation models.

Ablation Studies

TOSC Candidate Generation (TCG). TCG generates task-oriented shape completion candidates, which provide zero-shot generation capability to our method. To evaluate its impact, we conduct an ablation study by removing TCG and feeding the input partial point cloud to the TOSC restoration as the only candidate. As shown in Table 3, Results indicate that this removal significantly degrades performance, confirming its necessity.

TOSC Selection and Restoration (TSR). TSR selects the plausible shape and restores the shape. To evaluate its effect, we directly select the candidate with the minimum chamfer distance to the input point cloud and remove the shape restoration process. The inferior performance demonstrates its significance.

Task-Oriented Shape Completion (TOSC). We replace the task-oriented shape completion with a generic shape completion (Wang et al. 2024). The resulting performance drop confirms our idea of task-oriented shape completion.

Token Masking. The token masking scheme of the 3D DAE is proposed to improve the robustness. We remove this component and evaluate the performance. This variant exhibits a performance drop, confirming its effectiveness.

Gradient Guidance. Gradient guidance constrains the model to generate grasping actions that satisfy semantic and geometric consistency while optimizing the velocity field. To evaluate its necessity, we replace it with a direct loss-based optimization. This variant underperformed the full method, confirming the superiority of gradient guidance.

Conclusion

We have studied task-oriented shape completion, a new task that focuses on completing the potential contact regions rather than the entire shape. To achieve this, we propose a method that first generates multiple task-oriented shape completion candidates and then selects the most plausible one as well as optimizes its geometry by learning a 3D discriminative autoencoder. This method achieved state-of-the-art performance on both task-oriented dexterous grasping and task-oriented shape completion. In particular, it produces high-quality results in challenging scenarios.

Acknowledgments

This work was supported by the Natural Science Foundation of Hunan Province (2023JJ20051), the Science and Technology Innovation Program of Hunan Province (2023RC3011), the Cornerstone Foundation of NUDT (JS24-03) and the National Natural Science Foundation of China (62472434).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Chang, A. X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Feng, Q.; Lema, D. S. M.; Malmir, M.; Li, H.; Feng, J.; Chen, Z.; and Knoll, A. 2024. Dexgangrasp: Dexterous generative adversarial grasping synthesis for task-oriented manipulation. In *2024 IEEE-RAS 23rd International Conference on Humanoid Robots (Humanoids)*, 918–925. IEEE.

Hasson, Y.; Varol, G.; Tzionas, D.; Kalevatykh, I.; Black, M. J.; Laptev, I.; and Schmid, C. 2019. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11807–11816.

Huang, S.; Wang, Z.; Li, P.; Jia, B.; Liu, T.; Zhu, Y.; Liang, W.; and Zhu, S.-C. 2023. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16750–16761.

Huang, T.; Yan, Z.; Zhao, Y.; and Lee, G. H. 2024. ComPC: Completing a 3D Point Cloud with 2D Diffusion Priors. In *The Thirteenth International Conference on Learning Representations*.

Iwase, S.; Irshad, M. Z.; Liu, K.; Guizilini, V.; Lee, R.; Ikeda, T.; Amma, A.; Nishiwaki, K.; Kitani, K.; Ambrus, R.; et al. 2025. ZeroGrasp: Zero-Shot Shape Reconstruction Enabled Robotic Grasping. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 17405–17415.

Jian, J.; Liu, X.; Chen, Z.; Li, M.; Liu, J.; and Hu, R. 2025. G-DexGrasp: Generalizable Dexterous Grasping Synthesis Via Part-Aware Prior Retrieval and Prior-Assisted Generation. *arXiv preprint arXiv:2503.19457*.

Jian, J.; Liu, X.; Li, M.; Hu, R.; and Liu, J. 2023. Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14713–14724.

Jiang, H.; Liu, S.; Wang, J.; and Wang, X. 2021. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11107–11116.

Kakogeorgiou, I.; Gidaris, S.; Psomas, B.; Avrithis, Y.; Bursuc, A.; Karantzalos, K.; and Komodakis, N. 2022. What to hide from your students: Attention-guided masked image modeling. In *European Conference on Computer Vision*, 300–318. Springer.

Kasten, Y.; Rahamim, O.; and Chechik, G. 2023. Point cloud completion with pretrained text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36: 12171–12191.

Katz, S.; Tal, A.; and Basri, R. 2007. Direct visibility of point sets. In *ACM SIGGRAPH 2007 papers*, 24–es.

Kim, Y. H.; Kim, S.; Lee, Y.; and Park, F. C. 2025. DreamGrasp: Zero-Shot 3D Multi-Object Reconstruction from Partial-View Images for Robotic Manipulation. *arXiv preprint arXiv:2507.05627*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Li, A.; Zhu, Z.; and Wei, M. 2025. GenPC: Zero-shot Point Cloud Completion via 3D Generative Priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 1308–1318.

Li, H.; Mao, W.; Deng, W.; Meng, C.; Fan, H.; Wang, T.; Osamu, Y.; Tan, P.; Wang, H.; and Deng, X. 2024a. Multigraspllm: A multimodal llm for multi-hand semantic guided grasp generation. *arXiv preprint arXiv:2412.08468*.

Li, K.; Wang, J.; Yang, L.; Lu, C.; and Dai, B. 2024b. Semgrasp: Semantic grasp generation via language aligned discretization. In *European Conference on Computer Vision*, 109–127. Springer.

Liu, M.; Zhu, Y.; Cai, H.; Han, S.; Ling, Z.; Porikli, F.; and Su, H. 2023. Partslip: Low-shot part segmentation for 3d point clouds via pretrained image-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21736–21746.

Liu, Z.; Hu, L.; Zhou, T.; Tang, Y.; and Cai, Z. 2025. Prevalence Overshadows Concerns? Understanding Chinese Users' Privacy Awareness and Expectations Towards LLM-Based Healthcare Consultation. In *2025 IEEE Symposium on Security and Privacy (SP)*, 2716–2734. IEEE.

Mirjalili, R.; Krawez, M.; Silenzi, S.; Blei, Y.; and Burgard, W. 2024. LAN-grasp: An effective approach to semantic object grasping using large language models. In *First workshop on vision-Language Models for navigation and manipulation at ICRA 2024*.

Nichol, A.; Jun, H.; Dhariwal, P.; Mishkin, P.; and Chen, M. 2022. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*.

Pang, Y.; Tay, E. H. F.; Yuan, L.; and Chen, Z. 2023. Masked autoencoders for 3d point cloud self-supervised learning. *World Scientific Annual Review of Artificial Intelligence*, 1: 2440001.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

She, Q.; Zhang, S.; Ye, Y.; Hu, R.; and Xu, K. 2024. Learning Cross-Hand Policies of High-DOF Reaching and Grasping. In *European Conference on Computer Vision*, 269–285. Springer.

Tzionas, D.; Ballan, L.; Srikantha, A.; Aponte, P.; Pollefeys, M.; and Gall, J. 2016. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision*, 118(2): 172–193.

Uy, M. A.; Pham, Q.-H.; Hua, B.-S.; Nguyen, T.; and Yeung, S.-K. 2019. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1588–1597.

Wang, H.; Liu, Q.; Yue, X.; Lasenby, J.; and Kusner, M. J. 2021. Unsupervised point cloud pre-training via occlusion completion. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9782–9792.

Wang, J.; Cui, Y.; Guo, D.; Li, J.; Liu, Q.; and Shen, C. 2024. Pointattn: You only need attention for point cloud completion. In *Proceedings of the AAAI Conference on artificial intelligence*, volume 38, 5472–5480.

Wang, R.; Zhang, J.; Chen, J.; Xu, Y.; Li, P.; Liu, T.; and Wang, H. 2022. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. *arXiv preprint arXiv:2210.02697*.

Wei, Y.-L.; Jiang, J.-J.; Xing, C.; Tan, X.-T.; Wu, X.-M.; Li, H.; Cutkosky, M.; and Zheng, W.-S. 2024. Grasp as you say: Language-guided dexterous grasp generation. *Advances in Neural Information Processing Systems*, 37: 46881–46907.

Wei, Y.-L.; Lin, M.; Lin, Y.; Jiang, J.-J.; Wu, X.-M.; Zeng, L.-A.; and Zheng, W.-S. 2025. Afforddexgrasp: Open-set language-guided dexterous grasp with generalizable-instructive affordance. *arXiv preprint arXiv:2503.07360*.

Wu, T.; Pan, L.; Zhang, J.; Wang, T.; Liu, Z.; and Lin, D. 2021. Balanced chamfer distance as a comprehensive metric for point cloud completion. *Advances in Neural Information Processing Systems*, 34: 29088–29100.

Wu, T.; Zhang, J.; Fu, X.; Wang, Y.; Ren, J.; Pan, L.; Wu, W.; Yang, L.; Wang, J.; Qian, C.; et al. 2023. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 803–814.

Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; and Xiao, J. 2015. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1912–1920.

Yan, H.; Li, Z.; Luo, K.; Lu, L.; and Tan, P. 2025. SymmCompletion: High-Fidelity and High-Consistency Point Cloud Completion with Symmetry Guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9094–9102.

Yang, L.; Li, K.; Zhan, X.; Wu, F.; Xu, A.; Liu, L.; and Lu, C. 2022. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20953–20962.

Yu, X.; Rao, Y.; Wang, Z.; Liu, Z.; Lu, J.; and Zhou, J. 2021. Pointr: Diverse point cloud completion with geometry-aware transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12498–12507.

Yuan, W.; Khot, T.; Held, D.; Mertz, C.; and Hebert, M. 2018. Pcn: Point completion network. In *2018 international conference on 3D vision (3DV)*, 728–737. IEEE.

Zha, Y.; Ji, H.; Li, J.; Li, R.; Dai, T.; Chen, B.; Wang, Z.; and Xia, S.-T. 2024. Towards compact 3d representations via point feature enhancement masked autoencoders. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 6962–6970.

Zhang, J.; Xu, W.; Yu, Z.; Xie, P.; Tang, T.; and Lu, C. 2024. DexTOG: Learning Task-Oriented Dexterous Grasp With Language Condition. *IEEE Robotics and Automation Letters*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhang, R.; Guo, Z.; Gao, P.; Fang, R.; Zhao, B.; Wang, D.; Qiao, Y.; and Li, H. 2022. Point-m2ae: multi-scale masked autoencoders for hierarchical point cloud pre-training. *Advances in neural information processing systems*, 35: 27061–27074.

Zhang, X.; Zhang, S.; and Yan, J. 2024. Pcp-mae: Learning to predict centers for point masked autoencoders. *Advances in Neural Information Processing Systems*, 37: 80303–80327.

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

Zhong, Y.; Huang, X.; Li, R.; Zhang, C.; Liang, Y.; Yang, Y.; and Chen, Y. 2025. Dexgraspvla: A vision-language-action framework towards general dexterous grasping. *arXiv preprint arXiv:2502.20900*.

Zhu, Z.; Chen, H.; He, X.; Wang, W.; Qin, J.; and Wei, M. 2023. Svdformer: Complementing point cloud via self-view augmentation and self-structure dual-generator. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14508–14518.