FFHFlow: Diverse and Uncertainty-Aware Dexterous Grasp Generation via Flow Variational Inference

Qian Feng[*†1,3], **Jianxiang Feng**[*2,3], **Zhaopeng Chen**[2,], **Rudolph Triebel**[4,5] **and Alois Knoll**[3]

Abstract:

Synthesizing diverse, uncertainty-aware grasps for multi-fingered hands from partial observations remains a critical challenge in robot learning. Prior generative methods struggle to model the intricate grasp distribution of dexterous hands and often fail to reason about shape uncertainty inherent in partial point clouds, leading to unreliable or overly conservative grasps. We propose FFHFlow, a flow-based variational framework that generates diverse multi-finger grasps while explicitly quantifying perceptual uncertainty in the partial point clouds. Our approach leverages a normalizing flow-based deep latent variable model to learn a hierarchical grasp manifold, overcoming the mode collapse and rigid prior limitations of conditional Variational Autoencoders (cVAEs). By exploiting the invertibility and exact likelihoods of flows, FFHFlow introspects shape uncertainty in partial observations and identifies novel object structures, enabling risk-aware grasp synthesis. To further enhance reliability, we integrate a discriminative grasp evaluator with the flow likelihoods, formulating an uncertainty-aware ranking strategy that prioritizes grasps robust to shape ambiguity. Extensive experiments in simulation and real-world setups demonstrate that FFHFlow outperforms state-of-the-art baselines (including diffusion models) in grasp diversity and success rate, while achieving run-time efficient sampling. We also showcase its practical value in cluttered and confined environments, where diversity-driven sampling excels by mitigating collisions. (Project Page: https://sites.google.com/view/ffhflow/home/)

Keywords: Dexterous Grasping, Normalizing Flows, Uncertainty-Awareness

1 Introduction

Performing diverse dexterous grasps on various objects is important to realize advanced human-like robotic manipulation. However, achieving such capability remains challenging due to the high dimensionality of the hand configuration space and the high variability in the *object shape*, not to mention the high perceptual uncertainty in the *incomplete shape* from the partial observation.

Previous efforts to address this problem have garnered significant interest in Conditional Variational Autoencoder(cVAE) [1, 2]. However, its performance is hampered by the issues of *mode collapse* [3, 4, 5] and the often-used *overly simple prior* [6, 7]. These severely limit its ability to model the highly complex and inherently multi-modal grasp distribution. The lack of diversity significantly restricts the manipulability of robots operating in the common cluttered or confined space (*e.g.,* a two-tier shelf in Figure 4a) in daily life. Recent advances in the application of diffusion models [8] can potentially mitigate this issue but are struggling to achieve satisfactory run-time efficiency. More severely, only few prior studies [9, 10] considered the necessity of having *introspective capabilities* against the perceptual uncertainty raised by partial observation. Unfortunately, such solutions come with a often slow shape completion module,

[*]Equal Contributions, author ordering decided via coin-tossing: {qian, jianxiang}.feng@tum.de. †work done while at Agile Robots SE. [1]Amigos Robots [2]Agile Robots SE [3]School of Information Computation and Technology, Technical University of Munich (TUM) [4]Institute of Robotics and Mechatronics, German Aerospace Center (DLR) [5]Department of Informatics, Karlsruhe Institute of Technology (KIT)

9th Conference on Robot Learning (CoRL 2025), Seoul, Korea.

not to mention the issue of Out-of-Distribution (OOD) objects with unknown object shapes uncovered by the training data, which easily cause unexpected consequences, *e.g.,* grasp failure.

In this work, we introduce a novel flow-based variational grasp generation model to overcome the aforementioned challenges (Figure 1). To this end, we first observed deficient generalization gain by simply using Conditional Normalizing Flows (cNF) to model the point cloud-conditioned grasp distribution due to suboptimal latent features. To address this, we devise *FFHFlow-lvm* based on a novel flow-based Deep Latent Variable Models (DLVMs) (Figure 2). In this model, we first replace the commonly used isotropic Gaussian with an expressive *input-dependent* NF prior [4, 7] (Prior Flow). Further, we construct an elastic non-Gaussian likelihood function [3] (Grasp Flow) with another NF within the framework of DLVMs. With this design, we aim to conquer the over-regularization induced by the simple *input-independent* prior and the restricted form of the likelihood function (*e.g.,* isotropic Gaussian) in cVAE.

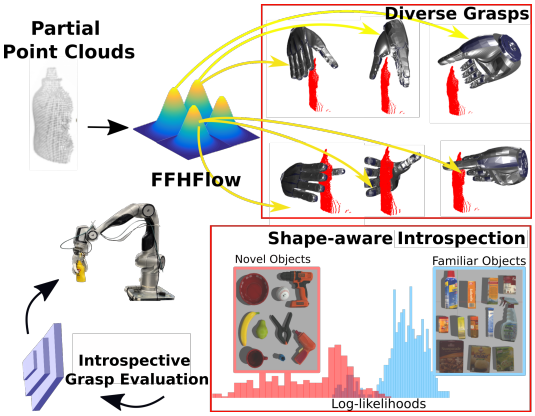


Figure 1: **Method overview**: in this work, we propose a variational grasp sampler based on Normalizing Flows (NFs), generating *diverse* dexterous grasps with shape-aware *introspection*.

To facilitate shape-aware introspection, inspired by [11], we exploit the exact likelihoods of NFs to represent the perceptual uncertainty (lower the more uncertain) related to the generated grasps. Specifically, the likelihoods of Grasp Flow account for the *incomplete shape* due to partial observability, namely **view uncertainty** (illustrated in Figure 5a). It predicts lower values for grasps toward the invisible views. Meanwhile, Prior Flow assigns lower likelihoods for OOD objects with *distinct shapes* to the objects in the training data, representing the **object uncertainty**. More noteworthy, the object uncertainty can be applied to promote robustness by detecting objects with novel geometry, which can easily result in grasp failures. Moreover, we leverage the view uncertainty to assist grasp evaluation and propose a new evaluation strategy based on a discriminative grasp evaluator.

To summarize, we contribute with **(1)** a novel flow-based DLVMs that can address the limitations of its alternatives, such as cVAE and diffusion models, **(2)** a new way to represent the perceptual uncertainties based on the flow likelihoods, leading to an uncertainty-aware grasp evaluation strategy and **(3)** a comprehensive experimental study both in simulation and on the real robot to verify the proposed idea, *i.e.,* object grasping from free space to clutter and confined space.

2 Related Work

Learning-based Grasp Synthesis. Deep generative models such as cVAEs [1, 12, 2, 13], auto-regressive models [14], cNFs [15, 16, 17], diffusion models [8, 18, 19] and Generative Adversarial Networks (GANs) [20, 21] are widely adopted in grasp synthesis with two-jaw gripper and multi-fingered hands. Among them, some [2, 14] address the partial observation challenge by employing time-intensive shape completion modules, which significantly slow down inference, reducing their practical applicability. Other approaches either assume the availability of complete object point clouds [16], a limitation in real-world scenarios, or only focus on a 2-jaw gripper [17]. Similarly, diffusion models [8, 18, 19] face runtime inefficiency due to their iterative denoising processes, even though there are some techniques [22, 23] to accelerate it. In contrast, our method, similar to [15, 12], eliminates the need for resource-intensive shape completion models by encoding partial observation information directly into latent variables. This lightweight approach allows our model to effectively handle partial observations while maintaining efficiency and diversity in grasp synthesis, enabling real-world applications such as grasping in clutter and dynamic grasping [24].

Uncertainty-Aware Grasping. Prior research on uncertainty-aware grasping with deep learning (DL) is primarily focused on suction cups [25, 26] and parallel-jaw grippers [27, 28, 29]. Their mo-

tivations and applications stem from the characteristic of uncertainty estimation in DL [30], ranging from efficient adaptation [27] and exploration [25] to grasp evaluation [28, 26, 29, 31, 32]. Although with encouraging results, their investigation is limited to a rather simplistic setting, i.e., either using simplified end-effectors or restricting the analysis to 2D spaces. In contrast, non-DL based uncertainty-aware grasping considers shape analysis [33, 34, 35] and multi-fingered hands [36, 37], which exhibits an understudied gap. Our work aims to bridge this gap by developing a holistic learning-based grasping model that can retrospectively reason about shape-aware uncertainty against partially observed and unknown objects. Some concurrent works [9, 10] share a similar spirit but rely on a time-consuming shape completion module, while our work is more real-time capable.

Normalizing Flows. Unlike other Deep Generative Models (DGMs), flow-based models [38, 39] can perform both exact likelihood evaluation and efficient sampling simultaneously. More noteworthy, NFs can be trained more stably when compared with GANs [40], perform better against the notorious *mode collapse* problem in both GANs and Variational Autoencoder (VAE) [5, 3, 4]. Moreover, we focus on discrete NFs, which is more run-time efficient than its continuous version based on flow matching [41] and does not require long trajectories in the de-nosing process like diffusion models. These appealing properties render NFs a promising tool for fast and effective probabilistic inference [42]. In addition, NFs found successful applications in point clouds processing [43, 44], feasibility learning [45], uncertainty estimation [46, 47], Out-of-distribution detection [48, 11]. Inspired by the probabilistic nature of NFs, we attempt to leverage this model to capture the multi-modal and complex grasp distribution and establish shape-aware introspective capability.

3 Preliminaries

Deep Latent Variable Models (DLVMs). In the context of modeling the unknown true data distribution $p^*(\mathbf{x})$ with a model $p_\theta(\mathbf{x})$ parameterized by θ based on a dataset $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$, latent variables $\mathbf{z}$ are usually introduced for discovering fine-grained factors controlling the data generating process or increasing the expressivity of the model $p_\theta(\mathbf{x})$. Latent variables $\{\mathbf{z}_i\}_{i=1}^N$ are part of the model but hidden and unobservable in the dataset. The resulting marginal probability is: $p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})d\mathbf{z}$. When $p_\theta(\mathbf{x}, \mathbf{z})$ is parameterized by Deep Neural Networks (DNNs), we term the model DLVMs [49, 3]. The difficulty of learning such models with Maximum Likelihood Estimation (MLE) lies in the intractability of the integral in the marginal probability for not having an analytic solution or efficient estimator. To remedy this, Variational Inference (VI) [50] provides a tractable lower bound of the marginal likelihood $p_\theta(\mathbf{x})$ to optimize by approximating the real posterior $p_\theta(\mathbf{z}|\mathbf{x})$ with an approximate one $q_\phi(\mathbf{z}|\mathbf{x})$:

$$\log p_\theta(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z}) + \log \frac{p_\theta(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}]. \tag{1}$$

When $q_\phi(\mathbf{z}|\mathbf{x})$ and $p_\theta(\mathbf{x}|\mathbf{z})$, are approximated by DNNs with an isotropic Gaussian as prior $p_\theta(\mathbf{z})$, we obtain the well-known instance of DLVMs, the VAE model [49].

Normalizing Flows. NFs are known to be universal distribution approximators [42]. That is, they can model any complex target distribution $p^*(\mathbf{x})$ on a space $\mathcal{R}^d$ by defining $\mathbf{x}$ as a transformation $T_\theta : \mathcal{R}^d \rightarrow \mathcal{R}^d$ parameterized by θ from a well-defined base distribution $p_u(\mathbf{u})$: $\mathbf{x} = T_\theta(\mathbf{u})$ where $\mathbf{u} \sim p_u(\mathbf{u})$, where $\mathbf{u} \in \mathcal{R}^d$ and p_u is commonly chosen as a unit Gaussian. By designing T_θ to be a *diffeomorphism*, that is, a bijection where both T_θ and T_θ^{-1} are differentiable, we can compute the likelihood of the input $\mathbf{x}$ exactly based on the change-of-variables formula [51]: $p_\theta(\mathbf{x}) = p_u(T_\theta^{-1}(\mathbf{x}))|\det(J_{T_\theta^{-1}}(\mathbf{x}))|$, where $J_{T_\theta^{-1}}(\mathbf{x}) \in \mathcal{R}^{d\times d}$ is the Jacobian of the inverse T_θ^{-1} with respect to $\mathbf{x}$. The transformation T_θ can be constructed by composing a series of bijective maps denoted by t_i, $T_\theta = t_1 \circ t_2 \circ ... \circ t_n$. When the target distribution is unknown, but samples thereof are available, we can estimate θ by minimizing the forward Kullback-Leibler Divergence (KLD), equivalent to minimizing the negative expected Log-Likelihood (LL):

$$\underset{\theta}{\operatorname{argmin}} \left[-\mathbb{E}_{p^*(\mathbf{x})}[\log(p_u(T_\theta^{-1}(\mathbf{x}))) + \log|\det(J_{T_\theta^{-1}}(\mathbf{x}))|] \right]. \tag{2}$$

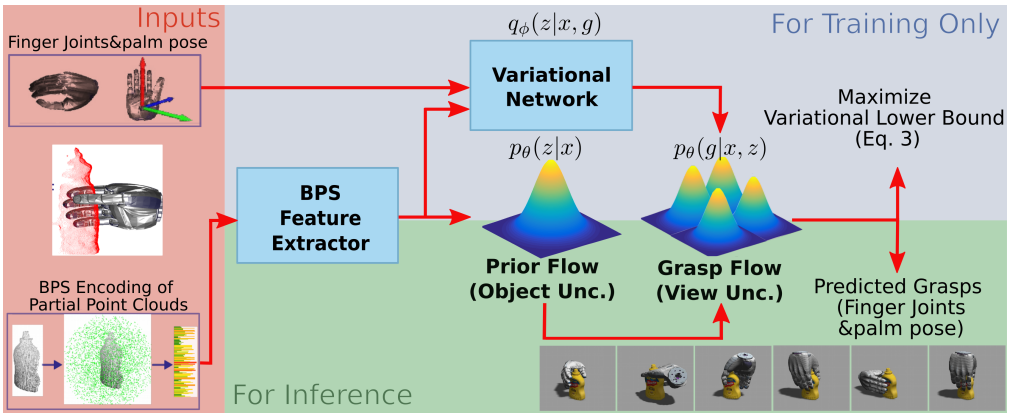


Figure 2: **Model Architecture**: at training time, the approximate posterior of the latent variable is inferred while the derived variational lower bound is maximized. During inference, Prior flow generates latent samples conditioned on the input point clouds, based on which Grasp Flow predicts the final grasps (best viewed in color).

4 Flow-based Grasp Synthesis

4.1 Problem Formulation

This work aims to generate diverse grasps given a *partial* point cloud denoted by $\mathbf{x} \in \mathbb{R}^{N\times 3}$. A grasp configuration $\mathbf{g} \in \mathbb{R}^d$ is represented by the 15-DOF hand joint configuration $\boldsymbol{j} \in \mathbb{R}^{15}$ and the 6D palm pose $(\mathbf{R}, \mathbf{t}) \in SE(3)$. Given an empirical dataset of N objects with N_i corresponding possible grasps $\mathcal{D} = \{\mathbf{x}_i, \{\mathbf{g}_{ik}\}_{k=1}^{N_i}\}_{i=1}^{N}$ drawn from the ground truth conditional distribution $p^*(\mathbf{g}|\mathbf{x})$, we train a probabilistic model $p_\theta(\mathbf{g}|\mathbf{x})$ parameterized by θ to approximate $p^*(\mathbf{g}|\mathbf{x})$. To facilitate shape-aware introspection [52], we exploit the log-likelihoods of $p_\theta(\mathbf{g}|\mathbf{x})$ to represent the uncertainty caused by incomplete point clouds and unknown object shapes.

4.2 Flow-based Grasp Sampler: FFHFlow-cnf

A straightforward idea to learn the conditional distribution $p_\theta(\mathbf{g}|\mathbf{x})$ is directly employing the cNF [53] without defining hidden variables on the latent space shown in Figure 3a. To this end, we condition the flow transformation T_θ and the base distribution p_u on the object point clouds $\mathbf{x}$, namely $T_{\theta|\mathbf{x}} : \mathcal{R}^d \times \mathcal{R}^l \rightarrow \mathcal{R}^d$, where l is the dimensionality of point cloud features and d for the grasp representation. We encode each point cloud with a fixed Basis Point Set (BPS) according to [54], resulting in a feature vector $\mathbf{x_b} \in \mathbb{R}^s$ of fixed length s, before being fed into the feature extractor network $f_\phi(\mathbf{x_b}) : \mathbb{R}^s \rightarrow \mathbb{R}^l$.

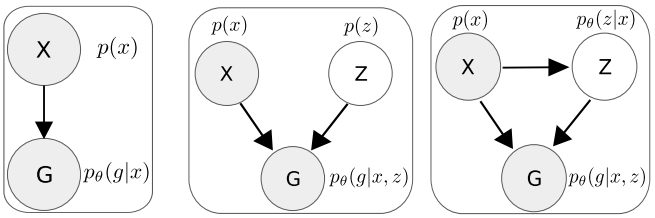
Limitation: Though *FFHFlow-cnf* achieved encouraging improvements in terms of diversity and accuracy when compared to the cVAE-based approach, we found it less generalizable with limited performance gain. We attribute the problem to the inadequate expressivity of the latent feature (explained in Section 3.3 of the appendix), especially when the model needs to understand the complicated relationships between the grasps and the partially observed point clouds of different objects. To address this problem, we introduce *FFHFlow-lvm* in the next sub-section, a flow-based variational sampler with a more expressive probabilistic representation in the latent space.

4.3 Flow-based Variational Grasp Sampler: FFHFlow-lvm

Inspired by the success of leveraging DLVMs for point cloud processing [43] and grasp generation [1, 12, 2], we devise a flow-based variational model that can induce expressive latent distribution for precise and diverse grasp generation. Specifically, we seek to overcome the over-regularization by the simplistic prior and the latent feature collapse by the Gaussian observation model in cVAE-based approaches [1, 12, 2]. This is achieved by introducing an input-dependent and expressive prior and a flexible observation model based on a cNF, which can be optimized efficiently under the framework of Stochastic Gradient Variational Bayes (SGVB) [55].

4.3.1 Learning Grasp Distribution via DLVMs

Our main idea is to introduce a latent variable into *FFHFlow-cnf* to increase the expressivity of the latent space. With the latent variable $\mathbf{z}$ shown in Figure 3c, we have the following conditional likelihoods of the grasps $\mathbf{g}$ given a partially observed point cloud $\mathbf{x}$. It can be factorized: $p_\theta(\mathbf{g}|\mathbf{x}) = \int p_\theta(\mathbf{g}|\mathbf{x},\mathbf{z})p_\theta(\mathbf{z}|\mathbf{x})d\mathbf{z}$. The likelihood is intractable due to the integral over the latent variables. We first derive a tractable lower bound for optimization (derivation in the appendix). Then, each component in the model is explained as follows how the pitfalls in cVAE are overcome. To note that θ denote the parameters for both flows and ϕ for the variational network.



(a) cNF (b) cVAE (c) FFHFlow-lvm

Figure 3: **Graphical illustration**. Shaded and unshaded circles denote observed and hidden variables, while arrows for dependencies. When compared to (a), DLVMs (b) and (c) have an extra hidden variable $\mathbf{z}$ for expressive latents learning. Though, the prior of $\mathbf{z}$ in cVAE (b) is an *input-independent* and simplistic Gaussian, while that of ours (c) is an *input-dependent* and more elastic cNF.

Variational Lower Bound: With Jensen Inequality, we have the following variational lower bound with an approximate posterior of the latent variable $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{g})$ (more details in the appendix):

$$\log p_\theta(\mathbf{g}|\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{g})}[\log p_\theta(\mathbf{g}|\mathbf{x},\mathbf{z})] - \beta KL(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{g})||p_\theta(\mathbf{z}|\mathbf{x})). \tag{3}$$

where β is a hyperparameter to control the extend of shape-aware information in the latents. In practice, the KLD term is implemented as the sum of the negative entropy of the approximate posterior and the cross-entropy between the prior and the approximate posterior for convenience. The effect of minimizing this term is to match the input-dependent prior $p_\theta(\mathbf{z}|\mathbf{x})$ to the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{g})$, and meanwhile encourage the approximate posterior to be more diverse. A flexible prior is of vital importance at inference time as we resort to the Prior Flow for drawing latent samples. These samples will be conditioned on the Grasp Flow for grasp generation (Figure 2). Therefore, *an input-dependent and learnable prior* is not only useful for increasing expressivity but also necessary for generating diverse grasps given an incomplete point cloud.

Grasp Flow $p_\theta(\mathbf{g}|\mathbf{x},\mathbf{z})$: Both VAE and GANs are reported to suffer from the *mode collapse* problem [5], which is also named *Information Preference Property* [4]. It illustrates the inability of VAE to capture the entire data distribution, *e.g.,* the complex multi-modal grasp distribution. Whenever this happens, the latent variable $\mathbf{z}$ is neglected by the powerful decoder and hence is *uninformative* in terms of the data, *i.e.,* grasps and object point clouds. It has been proved that the unbounded likelihood function is the crux of this problem, *e.g.,* an isotropic Gaussian [3]. To mitigate this, we propose to learn an cNF for the likelihood function $p_\theta(\mathbf{g}|\mathbf{x},\mathbf{z})$ by abuse of notations in Section 4.2: $p_\theta(\mathbf{g}|\mathbf{x},\mathbf{z}) = p_{u|\mathbf{z}}(T^{-1}_{\theta|\mathbf{z}}(\mathbf{g};\mathbf{z});\mathbf{z})|\det(J_{T^{-1}_{\theta|\mathbf{z}}}(\mathbf{g};\mathbf{z}))|$, where the base distribution $p_{u|\mathbf{z}} : \mathcal{R}^d \times \mathcal{R}^l \to \mathcal{R}$ is conditional on $\mathbf{z}$. For conciseness, $\mathbf{x}$ is not shown on the RHS as $\mathbf{z}$ already subsumes the information from $\mathbf{x}$, so Grasp Flow is simplified as $p_\theta(\mathbf{g}|\mathbf{z})$. In contrast to cVAE, there is no underlying assumption for the distribution form in NFs, which allows the model to learn a more general likelihood function instead of an isotropic Gaussian. Meanwhile, the architecture of NFs is restricted to be a diffeomorphism, which we anticipate to help alleviate the unboundness issue.

Prior Flow $p_\theta(\mathbf{z}|\mathbf{x})$: The overly-simple prior can induce excessive regularization, limiting the quality of the latent representation [6, 7], which can be observed in cVAE-based approaches [1, 12, 2] with an *input-independent* isotropic Gaussian as the prior (Figure 3b). On the other hand, the assumption of input-independence for the prior [4] poses difficulty in learning informative latent features. To address this, we propose to utilize a second cNF for an *input-dependent* prior distribution, in our case, a point cloud-dependent prior: $p_\theta(\mathbf{z}|\mathbf{x}) = p_u(T^{-1}_{\theta|\mathbf{x}}(\mathbf{z};\mathbf{x}))|\det(J_{T^{-1}_{\theta|\mathbf{x}}}(\mathbf{z};\mathbf{x}))|$.

Variational Network $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{g})$: The variational network is designed to approximate the real but intractable posterior distribution $p_\theta(\mathbf{z}|\mathbf{x},\mathbf{g})$ for amortized variational inference. $p_\theta(\mathbf{z}|\mathbf{x},\mathbf{g})$ is defined within the DLVM according to the Bayes formula: $p_\theta(\mathbf{z}|\mathbf{x},\mathbf{g}) = \frac{p_\theta(\mathbf{g}|\mathbf{z},\mathbf{x})p_\theta(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{g}|\mathbf{x})}$, where $p_\theta(\mathbf{g}|\mathbf{x}) =$

$\int p_\theta(\mathbf{g}|\mathbf{z},\mathbf{x})p_\theta(\mathbf{z}|\mathbf{x})d\mathbf{z}$ is the model evidence. From a pragmatic perspective, this network should be a powerful feature extractor for the grasps and object point clouds. As it is not used during inference, we keep it simple and use DNNs to predict a factorized Gaussian for the variational posterior distribution on the latent space $\mathbb{R}^l$, *i.e.,* $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{g}) = \mathcal{N}(\mathbf{z}; \mu_\phi(\mathbf{x},\mathbf{g}), diag(\delta_\phi(\mathbf{x},\mathbf{g})))$.

Grasp Generation and Shape-Aware Introspection: For a test object point cloud $\mathbf{x}^*$, we can generate the corresponding grasps $\mathbf{g}^*$ by performing ancestral sampling:

$$\mathbf{g}^* \sim p_\theta(\mathbf{g}|\mathbf{z}^*); \quad \mathbf{z}^* \sim p_\theta(\mathbf{z}|\mathbf{x}^*). \tag{4}$$

From Section 3 we know that we can compute the *exact likelihoods* of $p_\theta(\mathbf{g}|\mathbf{z}^*)$ and $p_\theta(\mathbf{z}|\mathbf{x}^*)$ from Grasp Flow and Prior Flow, respectively. We expect these two quantities to capture the knowledge gained by the model quantitatively. With these, our model can be introspective against its unknown knowledge, such as the *incomplete shape* due to partial observation and *unknown object shapes* due to limited coverage of training data. Specifically, the likelihoods of Grasp Flow quantify the inverse **view uncertainty** as it sees both the point cloud and the corresponding grasps through the learned latent variable. Thereby, high values are assigned to grasps towards visible views and low for grasps to the invisible views. Complementarily, with the proposed DLVMs, Prior Flow is expected to represent the inverse **object uncertainty** by purely conditioning on the input point cloud, *i.e.,* low likelihoods for unknown object shape and high for known ones.

4.4 Uncertainty-Aware Grasp Evaluation

To further secure the grasp success, we train a discriminative grasp evaluator $f_\psi(\mathbf{g},\mathbf{x})$ that outputs a score to better capture the grasp quality by learning to distinguish feasible grasps from infeasible ones in a supervised manner [12]. Furthermore, as Grasp Flow is able to quantify the view uncertainty, it is beneficial to incorporate this information into the grasp evaluation. With this, we can penalize the grasps approaching the uncertain side of the partially observed object to avoid potential collisions. To this end, we introduce an uncertainty-aware grasp evaluation strategy that fuses the batch-normalized view uncertainty into the grasp evaluator:

$$\epsilon f_\psi(\mathbf{g}^*,\mathbf{x}^*) + (1-\epsilon)\log p_\theta(\mathbf{g}^*|\mathbf{z}^*), \tag{5}$$

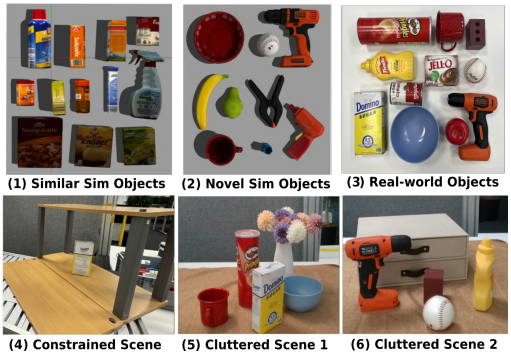
where $\epsilon \in [0,1]$ is a parameter for balancing the effects of grasp quality and reducing potential collisions due to the partial observability (an ablation study in Section 3.2 of the appendix).

5 Experiment

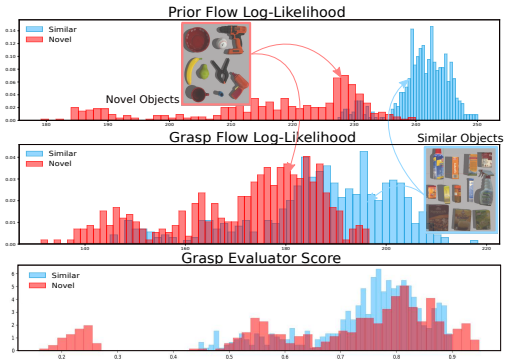
In this section, we perform comprehensive experiments and analysis to answer the following questions: **Q1**: Can the proposed *FFHFlow-lvm* facilitate more expressive latent representations for both diverse, high-quality and fast grasp synthesis? **Q2**: How well does *FFHFlow-lvm* perform compared with other state-of-the-art grasp synthesis approaches? **Q3**: How effective is the shape-aware introspection of *FFHFlow-lvm* against OOD data and does it benefit grasp evaluation? **Q4**: Can the model generalize to the complex real-world scenarios (*e.g.,* confined and cluttered scenes)? More ablation studies are included in appendix due to space limit.

5.1 Experimental Setup

The experiments of grasping single tabletop objects are performed in both the simulation and the real world with the DLR-HIT II hand [56]. In the real-world, except for single object grasping in unconfined space, we validate our approach in another two complex ones, *i.e.,* the confined and cluttered scenarios (Figure 4a). The success criterion is defined as the ability to lift the object 20 cm above its resting position without slippage. **Data:** We use only simulated data generated based on a heuristic grasp planner for training in Gazebo Simulator [12]. For **training**, we use 77 graspable objects filtered from KIT [57] datasets based on their graspability and object type. For **testing**, we select objects of two levels of difficulty from the KIT and YCB dataset [58] in simulation (Figure 4a): **(1) Similar**: 12 objects from KIT dataset with similar shapes to training objects, serving as **ID objects**. **(2) Novel**: 9 objects from YCB dataset with *shapes distinct from training objects*, often more difficult to grasp, serving as **OOD objects**. For real-world evaluation, we use 12 unknown objects from YCB dataset. (More detail in the appendix.)



(a) **Objects and setups evaluated in the simulation and real world.** We test simulation objects with similar (1) and novel (2) shapes, real objects (3), and grasping in confined space (4) and cluttered scene (5-6).



(b) **Analysis of view and object uncertainty (Top and Middle) and grasp evaluator scores (Bottom)** for In-Distribution (Blue) and Out-of-Distribution Objects (Red). Y axis: frequency/counts, X axis: scores/likelihoods (averaged over predicted grasps for Prior Flow and grasp evaluator). Flow likelihoods can better detect novel objects with low likelihoods.

Figure 4: **Experiment Setup (a) and Object Uncertainty Evaluation (b).**

Table 1: Results in Simulation (a) and Ablation Study on Uncertainty-aware Grasping (b).

(a) Average Success Rate and Run-time in Simulation

Methods	Similar	Novel	Run time (ms)
Heuristic	20.9%	11.1%	3387
cVAE [12]	84.6%	52.4%	**30**
GAN [21]	86.0%	49.4%	**30**
Diffusion [8]	88.2%	51.7%	1610
FFHFlow-cnf	85.4%	36.7%	70
FFHFlow-lvm	**94.6%**	**52.7%**	130

(b) Results of Uncertainty-aware Grasping

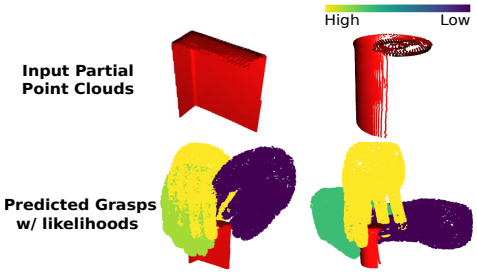
Evaluation Strategy	Similar	Novel
w/o Evaluator	41.4%	17.8%
+ Evaluator	90.5%	50.9%
Evaluator + Prior Flow	87.7%	52.4%
Evaluator + Grasp Flow	**94.6%**	**52.7%**

5.2 Evaluation in the Simulation

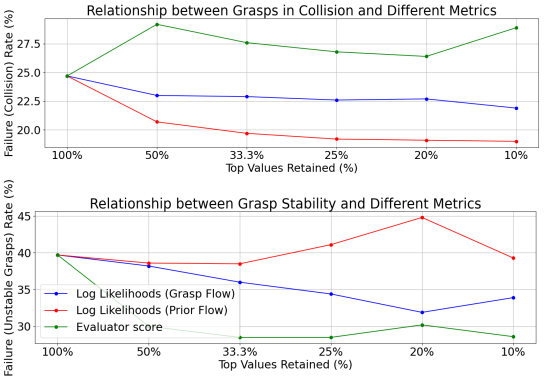
We demonstrate the simulation results through the success rate and runtime in Table 1a. **Baselines:** 1. Heuristic grasp sampler: a heuristic grasp sampler to generate grasps based on the normal of object point clouds [12]; 2. A cVAE-based approach, FFHNet [12]; 3. A GANs-based approach, DexGanGrasp [21]; 4. A Diffusion-based approach, DexDiffuser [8]. We apply the same evaluator for all generative grasping baselines except for *FFHFlow-lvm* with the proposed uncertainty-aware evaluation strategy. **Results:** To answer **Q1** and **Q2**, in Table 1a, first there is a clear performance drop from similar to novel objects, indicating the necessity of having the introspective capability to identify OOD objects. *FFHFlow-lvm* is able to generate high-quality and relatively fast grasp synthesis, outperforming the baselines on both similar objects and novel objects and is 10x faster than the diffusion baseline. Notably, the diffusion-based approach [8] demonstrates relatively better performance compared to cVAE [12] on similar objects and GAN [21] on both similar and novel objects. However, the iterative de-noising process results in a significantly longer runtime (1610ms).

5.3 Uncertainty-Aware Grasp Evaluation

Uncertainty Quantification: To answer **Q3**, Figure 4b shows that Prior Flow can better represent the **object uncertainty** via estimating the density of object shapes. It can better distinguish between in-distribution (ID) and novel/OOD objects with distinct shapes to the training data. This is useful in situations with multiple objects, where the robot could prioritize grasping in-distribution objects first to *avoid grasp failures*. On the other hand, Grasp Flow can represent the **view uncertainty** (visualized in Figure 5a). To showcase its effects on the predicted grasps, we present the relation between the number of collided and unstable grasps and increasing thresholds of the likelihoods and the grasp evaluator scores (higher the better) in Figure 5b. We assess the collision of the generated grasps with the Flexible Collision Library (FCL) and the grasp stability with Gazebo simulation (more implementation details in the appendix). We observe that, in contrast to the grasp evaluator



(a) **Visualizing view uncertainty**. Predicted grasps colored based on Grasp Flow likelihoods: lower for grasps (purple) to the *incomplete* face of the partially perceived point cloud.



(b) **Number of collided (Top) and unstable (Bottom) grasps** filtered with an increasing threshold (higher the better). Likelihoods from Grasp Flow (Blue) achieve a more optimal balance between grasp stability and collision.

Figure 5: **View Uncertainty Visualization (a) and Evaluation (b).**

score, both Prior Flow and Grasp Flow demonstrate the ability to reduce collision in the top plot. In the bottom plot, only the evaluator score and Grasp Flow can help improve grasp stability while the Prior Flow is less effective, which also explains the performance drop in Table 1b.

Uncertainty-Aware Grasp Evaluation: Grasp Flow likelihoods are utilized for grasp evaluation as it represents the view uncertainty and exhibits a better trade-off between grasp stability and collision avoidance in Figure 5b. This capability facilitates grasp selection with higher stability and less collision, leading to greater performance gain (Table 1b). The experiment in Table 1b follows the same set up as Table 1a and more details are described in Appendix Section 2.4.

5.4 Evaluation in the Real World

Table 2: Real-World Experiment Results

Workspace	Unconfined	Confined	Cluttered	
Metrics	Avg Succ Rate	Avg Succ Rate	Avg Succ Rate	Clearance Rate
cVAE [12]	62.5%	10.0%	68.2%	50.0%
FFHFlow-lvm	**77.5%**	**65.0%**	**76.0%**	**75.0%**

Unconfined Space: This setup is similar to the simulation, where 8 objects are grasped 10 times each. For **Q4**, Table 2 shows that *FFHFlow-lvm* is able to generalize to the real robot with a smaller gap. **Confined Space:** We selected a confined space, *i.e.,* a two-tier shelf to further mimic the realistic scenarios (d) in Figure 4a. In Table 2, a large performance gain is achieved by *FFHFlow-lvm* when compared to cVAE. The main failure of cVAE is due to its biased top grasps colliding with the environment (shelf). In such a confined space, a *diverse* grasp synthesizer is more effective than the mode-seeking cVAE. **Cluttered Scenes:** We conduct grasping experiments in cluttered scenes as (e) and (f) shown in Figure 4a. The metrics Clearance Rate (CR) denotes the probability of robots clear the scene. The predicted grasps with collision are filtered out based on FCL. *FFHFlow-lvm* achieves a higher success rate (SR) and also clearance rate (CR). More details in the appendix.

6 Conclusion

We introduce a novel flow-based variational approach, *FFHFlow-lvm* for generative grasp synthesis with better quality and diversity. This is achieved by mitigating the insufficiently informative latent features when applying cNF directly and overcoming problems in cVAE-based approaches, *i.e.,* , mode-collapse and the mis-specified prior, as well as inefficiency issues from diffusion-based approaches. Moreover, the model is equipped with shape-aware introspection quantified by the exact flow likelihoods, which further facilitates a novel hybrid scheme for uncertainty-aware grasp evaluation. Comprehensive experiments in the simulation and real world demonstrate strong performance and efficiency.

7 Limitations

Though our proposed idea exhibits numerous strengths, its limitations remain to solve for further improvement in the future. We highlight the most pronounced limitations to facilitate future research. **(1)** Trade-off between run-time and performance: reducing the flow size can improve the run-time but at expense of a slight performance loss. Investigating how to strike a balance is relevant for broader robotic applications. **(2)** The lack of abilities for efficient adaptation towards objects that differ significantly from those in the training dataset, which is hard to avoid for robots deployed in the wild [52, 59, 60]. We can further scale up the synthetic dataset, leveraging GPU-powered simulation such as Isaac Gym/Sim to increase generalization capability. **(3)** Sim-to-Real gap: Though the objects in simulation and real world are different in evaluation, a success rate drop of 17.1% encourages us to further investigate how to eliminate the sim-to-real gap, through various approaches such as camera noise modeling, physical parameters randomization and better simulators.

References

[1] A. Mousavian, C. Eppner, and D. Fox. 6-dof graspnet: Variational grasp generation for object manipulation. In *ICCV*, 2019.

[2] W. Wei, D. Li, P. Wang, Y. Li, W. Li, Y. Luo, and J. Zhong. Dvgg: Deep variational grasp generation for dextrous manipulation. *IEEE Robotics and Automation Letters*, 7(2):1659–1666, 2022.

[3] P.-A. Mattei and J. Frellsen. Leveraging the exact likelihood of deep latent variable models. *NeurIPS*, 31, 2018.

[4] S. Zhao, J. Song, and S. Ermon. Infovae: Balancing learning and inference in variational autoencoders. In *AAAI conference on artificial intelligence*, volume 33, pages 5885–5892, 2019.

[5] E. Richardson and Y. Weiss. On gans and gmms. *NeurIPS*, 31, 2018.

[6] F. P. Casale, A. Dalca, L. Saglietti, J. Listgarten, and N. Fusi. Gaussian process prior variational autoencoders. *NeurIPS*, 31, 2018.

[7] J. Tomczak and M. Welling. Vae with a vampprior. In *International Conference on Artificial Intelligence and Statistics*, pages 1214–1223. PMLR, 2018.

[8] Z. Weng, H. Lu, D. Kragic, and J. Lundell. Dexdiffuser: Generating dexterous grasps with diffusion models, 2024.

[9] J. Lundell, F. Verdoja, and V. Kyrki. Robust grasp planning over uncertain shape completions. In *2019 IEEE/RSJ IROS*, pages 1526–1532. IEEE, 2019.

[10] M. Humt, D. Winkelbauer, and U. Hillenbrand. Shape completion with prediction of uncertain regions. In *2023 IEEE/RSJ IROS*, pages 1215–1221. IEEE, 2023.

[11] J. Feng, J. Lee, S. Geisler, S. Günnemann, and R. Triebel. Topology-matching normalizing flows for out-of-distribution detection in robot learning. In *CoRL*, pages 3214–3241. PMLR, 2023.

[12] V. Mayer, Q. Feng, J. Deng, Y. Shi, Z. Chen, and A. Knoll. Ffhnet: Generating multi-fingered robotic grasps for unknown objects in real-time. In *2022 ICRA*, pages 762–769, 2022. doi: 10.1109/ICRA46639.2022.9811666.

[13] L. Zhang, K. Bai, G. Huang, Z. Bing, Z. Chen, A. Knoll, and J. Zhang. Multi-fingered robotic hand grasping in cluttered environments through hand-object contact semantic mapping, 2024.

[14] D. Winkelbauer, B. Bäuml, M. Humt, N. Thuerey, and R. Triebel. A two-stage learning architecture that generates high-quality grasps for a multi-fingered hand. In *2022 IEEE/RSJ IROS*, pages 4757–4764, 2022. doi:10.1109/IROS47612.2022.9981133.

[15] M. Yan, A. Li, M. Kalakrishnan, and P. Pastor. Learning probabilistic multi-modal actor models for vision-based robotic grasping. In *2019 ICRA*, pages 4804–4810. IEEE, 2019.

[16] Y. Xu, W. Wan, J. Zhang, H. Liu, Z. Shan, H. Shen, R. Wang, H. Geng, Y. Weng, J. Chen, et al. Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In *Proceedings of the IEEE/CVF CVPR*, pages 4737–4746, 2023.

[17] B. Lim, J. Kim, J. Kim, Y. Lee, and F. C. Park. Equigraspflow: SE(3)-equivariant 6-dof grasp pose generative flows. In *CoRL*, 2024.

[18] J. Zhang, H. Liu, D. Li, X. Yu, H. Geng, Y. Ding, J. Chen, and H. Wang. Dexgraspnet 2.0: Learning generative dexterous grasping in large-scale synthetic cluttered scenes, 2024.

[19] J. Carvalho, A. T. Le, P. Jahr, Q. Sun, J. Urain, D. Koert, and J. Peters. Grasp diffusion network: Learning grasp generators from partial point clouds with diffusion models in so(3)xr3, 2024. URL `https://arxiv.org/abs/2412.08398`.

[20] F. Patzelt, R. Haschke, and H. J. Ritter. Conditional wgan for grasp generation. In *ESANN*, 2019.

[21] Q. Feng, D. S. M. Lema, M. Malmir, H. Li, J. Feng, Z. Chen, and A. Knoll. Dexgangrasp: Dexterous generative adversarial grasping synthesis for task-oriented manipulation, 2024.

[22] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[23] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models, 2022.

[24] Y. Burkhardt, Q. Feng, J. Feng, K. Sharma, Z. Chen, and A. Knoll. Multi-fingered dynamic grasping for unknown objects, 2024.

[25] Y. Shi, P. Schillinger, M. Gabriel, A. Qualmann, Z. Feldman, H. Ziesche, and N. A. Vien. Uncertainty-driven exploration strategies for online grasp learning. In *2024 IEEE ICRA*, pages 781–787, 2024. doi:10.1109/ICRA57147.2024.10610056.

[26] R. Cao, B. Yang, Y. Li, C.-W. Fu, P.-A. Heng, and Y.-H. Liu. Uncertainty-aware suction grasping for cluttered scenes. *IEEE Robotics and Automation Letters*, 9(6):4934–4941, 2024. doi:10.1109/LRA.2024.3385609.

[27] H. Zhu, Y. Li, F. Bai, W. Chen, X. Li, J. Ma, C. S. Teo, P. Yuen Tao, and W. Lin. Grasping detection network with uncertainty estimation for confidence-driven semi-supervised domain adaptation. In *2020 IEEE/RSJ IROS*, pages 9608–9613, 2020. doi:10.1109/IROS45743.2020.9341056.

[28] B. Stephan, D. Aganian, L. Hinneburg, M. Eisenbach, S. Müller, and H.-M. Gross. On the importance of label encoding and uncertainty estimation for robotic grasp detection. In *2022 IEEE/RSJ IROS*, pages 4781–4788. IEEE, 2022.

[29] Y. Shi, E. Welte, M. Gilles, and R. Rayyes. vmf-contact: Uncertainty-aware evidential learning for probabilistic contact-grasp in noisy clutter. *arXiv preprint arXiv:2411.03591*, 2024.

[30] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023.

[31] T. G. W. Lum, A. H. Li, P. Culbertson, K. Srinivasan, A. D. Ames, M. Schwager, and J. Bohg. Get a grip: Multi-finger grasp evaluation at scale enables robust sim-to-real transfer, 2024. URL https://arxiv.org/abs/2410.23701.

[32] Z. Zheng, Q. Feng, H. Li, A. Knoll, and J. Feng. Evaluating uncertainty-based failure detection for closed-loop llm planners. *arXiv e-prints*, pages arXiv–2406, 2024.

[33] V. N. Christopoulos and P. Schrater. Handling shape and contact location uncertainty in grasping two-dimensional planar objects. In *2007 IEEE/RSJ IROS*, pages 1557–1563, 2007. doi: 10.1109/IROS.2007.4399509.

[34] S. Dragiev, M. Toussaint, and M. Gienger. Gaussian process implicit surfaces for shape estimation and grasping. In *2011 IEEE ICRA*, pages 2845–2850, 2011. doi:10.1109/ICRA.2011.5980395.

[35] D. Chen, V. Dietrich, Z. Liu, and G. Von Wichert. A probabilistic framework for uncertainty-aware high-accuracy precision grasping of unknown objects. *Journal of Intelligent & Robotic Systems*, 90:19–43, 2018.

[36] M. Li, K. Hang, D. Kragic, and A. Billard. Dexterous grasping under shape uncertainty. *Robotics and Autonomous Systems*, 75:352–364, 2016.

[37] S. Chen, J. Bohg, and C. K. Liu. Springgrasp: An optimization pipeline for robust and compliant dexterous pre-grasp synthesis. *arXiv preprint arXiv:2404.13532*, 2024.

[38] L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[39] D. P. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *NeurIPS*, 31, 2018.

[40] L. Mescheder, A. Geiger, and S. Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR, 2018.

[41] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling, 2023.

[42] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, 22(1):2617–2680, 2021.

[43] R. Klokov, E. Boyer, and J. Verbeek. Discrete point flow networks for efficient point cloud generation. In *ECCV*, pages 694–710. Springer, 2020.

[44] J. Postels, M. Liu, R. Spezialetti, L. Van Gool, and F. Tombari. Go with the flows: Mixtures of normalizing flows for point cloud generation and reconstruction. In *3DV*, pages 1249–1258. IEEE, 2021.

[45] J. Feng, M. Atad, I. V. Rodriguez Brena, M. Durner, and R. Triebel. Density-based feasibility learning with normalizing flows for introspective robotic assembly. In *18th RSS 2023 Workshops*, 2023.

[46] B. Charpentier, D. Zügner, and S. Günnemann. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. *NeurIPS*, 33:1356–1367, 2020.

[47] J. Postels, H. Blum, Y. Strümpler, C. Cadena, R. Siegwart, L. Van Gool, and F. Tombari. The hidden uncertainty in a neural networks activations. *arXiv preprint arXiv:2012.03082*, 2020.

[48] P. Kirichenko, P. Izmailov, and A. G. Wilson. Why normalizing flows fail to detect out-of-distribution data. *NeurIPS*, 33:20578–20589, 2020.

[49] D. P. Kingma, M. Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

[50] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[51] V. I. Bogachev and M. A. S. Ruas. *Measure theory*. Springer, 2007.

[52] J. Feng, M. Durner, Z.-C. Márton, F. Bálint-Benczédi, and R. Triebel. Introspective robot perception using smoothed predictions from bayesian neural networks. In *Robotics Research*, pages 660–675, Cham, 2019. Springer International Publishing. ISBN 978-3-030-95459-8.

[53] C. Winkler, D. Worrall, E. Hoogeboom, and M. Welling. Learning likelihoods with conditional normalizing flows. *arXiv preprint arXiv:1912.00042*, 2019.

[54] S. Prokudin, C. Lassner, and J. Romero. Efficient learning on point clouds with basis point sets. In *ICCV*, pages 4332–4341, 2019.

[55] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[56] H. Liu, K. Wu, P. Meusel, N. Seitz, G. Hirzinger, M. Jin, Y. Liu, S. Fan, T. Lan, and Z. Chen. Multisensory five-finger dexterous hand: The dlr/hit hand ii. In *2008 IEEE/RSJ IROS*, pages 3692–3697. IEEE, 2008.

[57] A. Kasper, Z. Xue, and R. Dillmann. The kit object models database: An object model database for object recognition, localization and manipulation in service robotics. *IJRR*, pages 927–934, 2012.

[58] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *ICAR*, pages 510–517. IEEE, 2015.

[59] J. Feng, J. Lee, M. Durner, and R. Triebel. Bayesian active learning for sim-to-real robotic perception. In *2022 IEEE/RSJ IROS*, pages 10820–10827. IEEE, 2022.

[60] M. Noseworthy, S. Seiji, C. Kessens, and N. Roy. Amortized inference for efficient grasp model adaptation. In *ICRA*. IEEE, 2024.

Supplementary Materials for "FFHFlow: Diverse and Uncertainty-Aware Dexterous Grasp Generation via Flow Variational Inference"

Contents

1 Derivation of Variational Lower Bound

To learn a probabilistic model $p_\theta(\mathbf{g}|\mathbf{x})$ parameterized by θ, we optimize it by maximizing its variational lower bound. Here $\mathbf{g}$ denotes grasp configuration, and $\mathbf{x}$ is a partially observed point cloud. Assume that the real posterior $p_\theta(z|x,g)$ is defined within the deep lantent variation model framework according to the Bayes formula: $p_\theta(z|x,g) = \frac{p_\theta(g|z,x)p_\theta(z|x)}{p_\theta(g|x)}$, where $p_\theta(g|x) = \int p_\theta(g|z,x)p_\theta(z|x)dz$ is the so-called model evidence. Based on Jensen Inequality, we can derive the variational lower bound with an approximate posterior of the latent variable $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{g})$ step by step with the following:

$$
\begin{aligned}
\log p_\theta(\mathbf{g}|\mathbf{x}) &= \log \int p_\theta(\mathbf{g},\mathbf{z}|\mathbf{x})d\mathbf{z} \\
&= \log \int p_\theta(\mathbf{g}|\mathbf{z},\mathbf{x})p(\mathbf{z}|\mathbf{x})d\mathbf{z} \\
&= \log \int p_\theta(\mathbf{g}|\mathbf{z},\mathbf{x})\frac{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{g})}{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{g})}p_\theta(\mathbf{z}|\mathbf{x})d\mathbf{z} \\
&\geq \int q_\phi(\mathbf{z}|\mathbf{x},\mathbf{g}) \log [\frac{p_\theta(\mathbf{g}|\mathbf{z},\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{g})}p_\theta(\mathbf{z}|\mathbf{x})]d\mathbf{z} \\
&= \int q_\phi(\mathbf{z}|\mathbf{x},\mathbf{g}) \log p_\theta(\mathbf{g}|\mathbf{z},\mathbf{x})d\mathbf{z} - \\
&\quad \int q_\phi(\mathbf{z}|\mathbf{x},\mathbf{g}) \log \frac{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{g})}{p_\theta(\mathbf{z}|\mathbf{x})}d\mathbf{z} \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{g})}[\log p_\theta(\mathbf{g}|\mathbf{z},\mathbf{x})] - KL(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{g})||p_\theta(\mathbf{z}|\mathbf{x})).
\end{aligned}
\tag{1}
$$

Here $p_\theta(\mathbf{g}|\mathbf{z},\mathbf{x})$ represents our Grasp Flow, $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{g})$ for Variational Network, and $p_\theta(\mathbf{z}|\mathbf{x})$ for Prior Flow. In practice, we add a hyper-parameter β for the second KL-divergence term to control the trade-off between the information stored in the latent and the regularization induced by the structure of the prior. Concisely speaking, a high β will enable the model learn more shape-aware lantents but may be less informative for grasp prediction and vice versus. Therefore, a proper value is to be tuned for achieving a balance according to the specific task.

2 Data Generation Pipeline

2.1 Training, Evaluation and Testing Objects

For the data generation, we collect 89 graspable objects filtered from KIT datasets according to their graspability. We split 89 KIT [1] objects into a training set containing 77 objects, shown in Figure 1, and a test set of 12 objects as "similar" objects (Baking Soda, Bath Detergent, Broccoli Soup, Cough Drops, Curry, Fizzy Tablets, Instant Sauce, Nut Candy, Potato Dumpling, Spray Flask, Tomato Soup, Yellow SaltCube). We further include 9 YCB objects with distinct geometric shapes to the training set as "novel" objects (Bowl, Baseball, Power Drill, Plastic Pear, Plastic Banana, Mug, Clamp, Toy Airplane parts), illustrated in Figure 2.

2.2 Heuristic grasp planner

We derive the heuristic grasp planner from [2]. Here is a more detailed explanation. The heuristic grasp planner samples the grasp poses based on the normal of each object point cloud. We extend the target point cloud in the normal direction with a random value between 4.5 and 11.5 cm. Then, we add translation noise of $\pm 1cm$ in 3D space. To improve the data generation efficiency, the y-axis of the palm pose is aligned with the more extended object side and oriented upwards. Afterward, we add rotation noise of ± 0.7 rad around the x-direction and ± 0.35 rad around y- and z- directions.

To efficiently sample the 15-dof joint configuration, we apply eigengrasps from Ciocarlie [3] to sample a valuable subspace. Ciocarlie's work was inspired by the Neuroscience community, which showed that the joint DoFs of human hands during real-world grasping trials were primarily not operating independently but coordinated. More than 80 % of the variation in the data could be explained by the two first components of the principal component analysis (PCA). These components

Figure 1: **The training objects from KIT dataset for data generation**

Figure 2: **The testing objects of 12 KIT dataset as "similar" and 9 YCB objects as "novel".**

were termed eigengrasps, as almost any grasp joint configuration can be synthesized as a linear combination of a few eigengrasps. Thus, we design four eigengrasps $\mathbf{e_1}, \mathbf{e_2}, \mathbf{e_3}, \mathbf{e_4} \in \mathbb{R}^{15}$.

$$\theta = \sum_{i=1}^{4} k_i \mathbf{e}_i \tag{2}$$

The full joint configuration θ is computed through sampling the coefficients $k_i \in [0, 1]$.

2.3 Grasp data generation pipeline

The data generation pipeline is similar with [2]. First, we randomly spawn an object in front of the robot. Then, a point cloud is recorded by the simulated camera. Afterward, we generate grasp samples based on heuristic grasp planner explained in Section 2.2, which are further filtered by Moveit in terms of reachability and collision. The robot then will execute the sampled grasp and attempt to lift it, where the grasp success is labeled automatically. This process is repeated for all objects with multiple random poses.

Since the grasp distribution is only object-dependent, the model should predict the same grasp distribution given different partial views of the same object. Therefore, we apply a data augmentation strategy by randomly spawning every object with 50 different initial poses to increase the dataset capacity by 50 times. In total, we generated a dataset of around 180k grasps, of which 30k resulted in success.

2.4 Experiment setup

Figure 3 shows our simulation setup. We use a Panda robot model with the DLR-HIT II hand as the end-effector. A simulated Realsense D415 camera is used to capture the point cloud. Afterward, the scene point cloud is captured by a Realsense D415 camera and then segmented with plane removal from RANSAC [4] to obtain the segmented object point cloud. The Basis Point Set (BPS)-encoded point cloud, after being segmented with plane removal, is fed as input to different models to grasp synthesis and ranking. Grasping success is defined as the ability for the DLR-HIT Hand II to lift the object 20 cm above its resting position without slippage.

The top grasp with the highest score is subsequently selected for execution. We conduct up to 20 trials per object in our simulation experiments. To facilitate a fair comparison for the grasp generator without a grasp evaluator, we evaluate the grasp samplers in simulation by executing the top 20 grasps instead of the single top-most one.

We choose the 12 test objects from the KIT dataset for the experiments in simulation. Each test object is spawned in simulation 20 times in random positions and random yaw-angle orientations. After recording each point cloud, we segment the object from the ground plane via RANSAC [4]. We combine the segmented object point cloud with random samples from the base distribution of FFHFlow model, namely a univariate Gaussian $\{\mathbf{z}\}^{100} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to generate 100 grasps per object. Afterward, Grasp Evaluator will rank all the generated grasps with predicted success probability. The grasp with the highest score will be executed. Therefore, we include grasp failures, which happen during the grasp execution phase, but exclude the failures where the robot collides with the object on the way to reach the grasp pose.

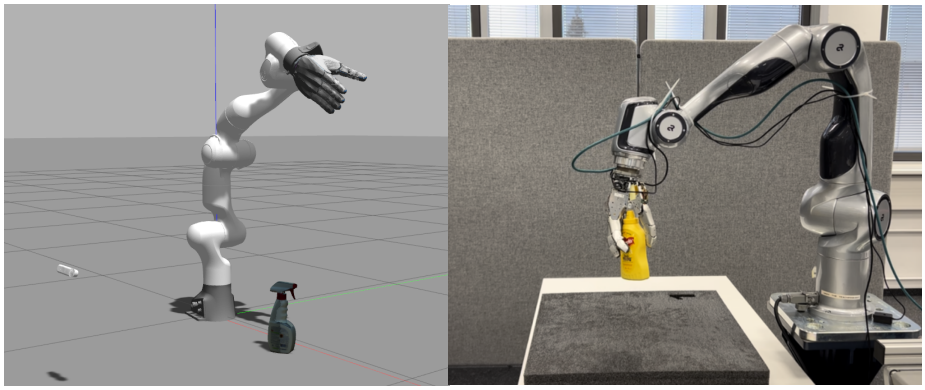


Figure 3: **The simulation setup in Gazebo and the real world setup**

The real-world setup is also demonstrated in Figure 3. For each method, we perform 80 grasps with 8 YCB objects [5] in the real world with a free workspace, and 20 grasps with 4 objects in a constrained workspace. Since our method is robot-independent, we choose Diana's robot arm, which is kinematically similar to the Franka robot, for real-world experiments.

2.5 Implementation Details

We first pre-process the input point cloud with BPS encoding [6], which is reported to work similarly well with PointNet++ [7] but with less compute. This reduces the overall computation and decrease the inference time. Afterwards the point cloud features are extracted with fully-connected residual block (FC Resblock) and are further conditioned on the flow model. We use a similar architecture for the point cloud feature extractor, variational inference network, and grasp evaluator, *i.e.*, a network with multiple fully-connected residual blocks. We use skip connections from each input to each fully-connected residual block (FC ResBlock) or fully-connected (FC) layer. The core building block of both models is the FC ResBlock, which consists of two parallel paths from input to output. One path consists of a single FC layer, the other path has two FC layers. Each is followed by a layer of batchnorm (BN).

Based on the *normflows* package [8], we implement the Grasp Flow (for both *FFHFlow-cnf* and *FFHFlow-lvm*) and Prior Flow (only in *FFHFlow-lvm*) with an 8-layered conditional Glow [9] where each layer has a 4-layered Multi-Layer Perceptron (MLP) for predicting the parameters of the affine operation. Both models are trained with a learning rate of $1\mathrm{e}{-4}$ and a mini-batch size of 64 for 16 epochs or $20k$ iterations. The objective in Equation (1) is then optimized with a linearly increased β from $1\mathrm{e}{-7}$ to $1\mathrm{e}{-1}$ in each iteration based on the *AdamW* optimizer [10] for *FFHFlow-lvm*. We also use Monte Carlo sampling to approximate the expectation operation in Equation (1). The number of samples is empirically set to 1. Moreover, during evaluation, we apply a positive offset of 0.2 rad on predicted joint configurations to ensure a more stable grasp.

2.6 Metric: Coverage

Coverage (Cov): It measures the fraction of grasps in the ground truth grasp set $\mathbf{G}_{gt}$ that is matched to at least one grasp in the generated set $\mathbf{G}_{gen}$:

$$Cov(\mathbf{G}_{gen}, \mathbf{G}_{gt}) = \frac{|\{\arg\min_{\mathbf{G}_{gt}} d(\mathbf{g}_{gen}, \mathbf{g}_{gt}) | \mathbf{g}_{gen} \in \mathbf{G}_{gen}\}|}{|\mathbf{G}_{gt}|} \tag{3}$$

For each grasp in the generated set $\mathbf{G}_{gen}$, its nearest neighbor based on L2 distance in the ground truth set $\mathbf{G}_{gt}$ is marked as a match. Coverage (Cov) can be used to quantify the diversity of the generated grasp set with the ground truth set as reference.

3 Additional Experimental Results

Table 1: Results Comparison on Cov and Run-time

Methods (w/o eval)	Cov ↑	Run-time ↓
FFHNet [2]	$22.5\% \pm 1.6\%$	30ms
FFHNet-prior	$24.4\% \pm 1.0\%$	31ms
FFHFlow-cnf	$30.0\% \pm 0.2\%$	70ms
FFHFlow-lvm	$30.3\% \pm 0.3\%$	130ms
FFHFlow-lvm-light	$29.9\% \pm 0.4\%$	60ms

3.1 Per-object Simulation and Real-world Results

We also include all the per-object simulation results in Table. 2 and Table. 3. We also include per-object results for real-world experiment with unconfined workspace in Table. 4 and with confined workspace in Table.5. Note that in Table.4, Chips Can have no results since they are too large for our hand to grasp.

Table 2: Per-object Success Rate Comparison for Similar Objects in Simulation

Methods	Objects												Average Succ Rate
	Baking Soda	Bath Detergent	Broccoli Soup	Cough Drops	Curry	Fizzy Tablets	Instant Sauce	Nut Candy	Potato Dumpling	Spray Flask	Tomato Soup	Yellow SaltCube	
Heuristic	6/20	2/20	4/20	4/20	8/20	8/20	2/20	4/20	3/20	2/20	3/20	4/20	20.9%
cVAE [2]	19/20	**19/20**	**19/20**	18/20	**19/20**	**20/20**	15/20	12/20	16/20	19/20	13/20	14/20	84.6%
GAN [11]	16/20	15/20	17/19	18/19	**19/20**	15/19	18/20	14/20	13/20	**19/19**	**20/20**	**19/20**	86.0%
Diffusion [12]	19/20	14/20	16/20	18/19	18/20	19/20	16/20	17/20	**20/20**	17/19	19/20	18/20	88.2%
FFHFlow-cnf	19/20	18/20	17/20	17/20	17/20	**20/20**	15/20	15/20	17/20	18/20	15/20	17/20	85.4%
FFHFlow-lvm	**20/20**	**19/20**	**19/20**	**19/20**	**19/20**	**20/20**	**20/20**	**18/20**	18/20	**20/20**	17/20	18/20	**94.6%**

Table 3: Per-object Success Rate Comparison for Novel Objects in Simulation

Methods	Objects									Average Succ Rate
	Power Drill	Baseball	Bowl	Mug	Pear	Banana	Extra Large Clamp	C Toy Airplane	B Toy Airplane	
Heuristic	3/20	2/20	4/20	9/20	3/20	3/20	0/20	8/20	0/20	17.8%
cVAE [2]	12/20	16/20	3/10	13/20	12/20	5/20	**5/20**	**19/20**	4/20	52.4%
GAN [11]	**15/18**	**17/20**	**7/12**	7/12	14/20	1/20	1/20	11/20	**7/20**	49.4%
Diffusion [12]	12/19	15/20	7/17	13/20	16/20	**10/20**	0/20	15/20	3/20	51.7%
FFHFlow-cnf	13/20	14/20	2/17	**15/20**	8/20	1/20	1/20	9/20	2/20	36.7%
FFHFlow-lvm	13/18	13/20	2/11	14/20	**18/20**	6/20	3/20	15/20	5/20	**52.7%**

3.2 Uncertainty-aware Grasp Evaluation

Uncertainty Quantification: For the experiment conducted for Figure 4, we collect an evaluation set and generate 100 grasp candidates for each partial view. For each grasp, we obtain the likelihoods of Grasp Flow and Prior Flow, as well as the evaluator scores. To assess the quality of the generated grasps, we utilize the Flexible Collision Library (FCL) to predict collisions for each grasp and Gazebo to evaluate the grasp stability of the remaining non-collided grasps. The x-axis represents the percentage of top-ranked values retained, ranging from 100% to 10%, while the y-axis shows the failure rate.

In Figure 4, we observe a clear *negative* correlation between the grasp evaluator score and the failure rate due to collision. In contrast, Prior Flow and Grasp Flow demonstrate the ability to reduce collision. among which Prior Flow exhibits the strongest correlation with the collision rate, highlighting its potential for capturing shape awareness. In the bottom plot, both the evaluator score and Grasp Flow likelihoods exhibit a strong correlation with grasp stability. The grasp evaluator outperforms Grasp Flow as it was specifically trained to distinguish positive grasps from negative ones. However, the Prior Flow, representing the object-level shape uncertainty, is less relevant to grasp stability.

Ablation Study: We conduct an ablation study presented in Table 6 to understand the trade-off between increasing grasp quality (grasp evaluator) and lowering view-level shape uncertainty (Grasp Flow), namely the optimal value of the additive coefficient (ϵ). By increasing the impact of lowering ϵ, we can see the performance first increases and drops. The optimal value is 0.01, indicating the major contribution to grasp success from the grasp evaluator.

Table 6: ϵ in Introspective Grasp Evaluation

Additive Coefficients (ϵ)	0.0	0.01	0.1	0.5	1.0
Similar	90.5%	**94.6%**	90.6%	78.6%	63.0%
Novel	50.9%	**52.7%**	50.9%	34.3%	25.3%

Table 4: Per-object Success Rate Comparison for Objects in Real-World Unconfined Workspace

Methods	Objects									Average Succ Rate
	Sugar Box	Apple	Tomate Soup Can	Pudding Box	Mug	Mustard Bottle	Chips Can	Baseball	Foam Brick	
cVAE [2]	**9/10**	2/10	6/10	**10/10**	3/10	6/10	-	4/10	**10/10**	62.5%
FFHFlow-cnf	4/10	4/10	**8/10**	9/10	**7/10**	4/10	-	5/10	**10/10**	63.75%
FFHFlow-lvm	8/10	**6/10**	**8/10**	**10/10**	6/10	**7/10**	-	**8/10**	9/10	**77.5%**

Table 5: Per-object Success Rate Comparison for Objects in Real-World Confined Workspace

Methods	Objects				Average Succ Rate
	Foam Brick	Pudding Box	Baseball	Tomato Soup Can	
cVAE [2]	0/5	0/5	0/5	2/5	10.0%
FFHFlow-lvm	4/5	3/5	3/5	3/5	**65.0%**

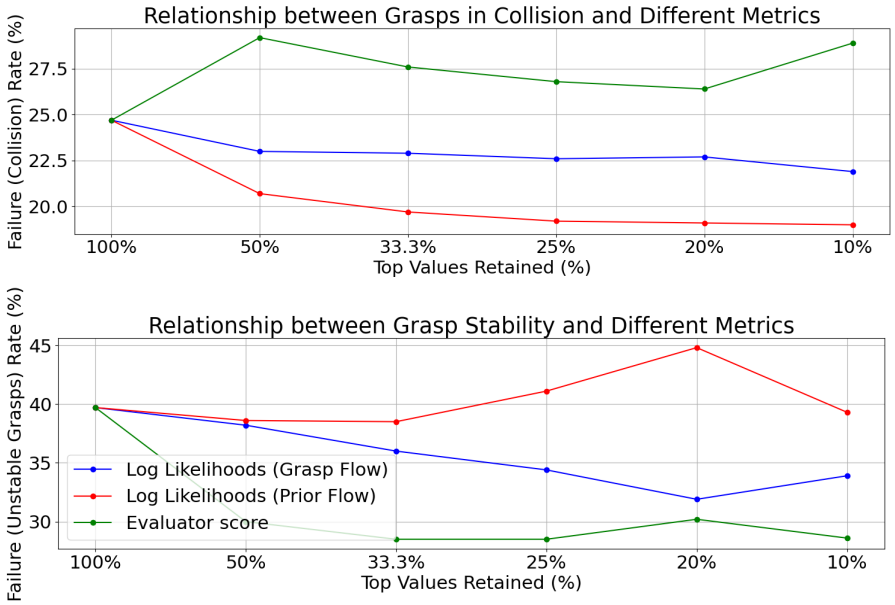


Figure 4: **Number of collided (Top) and unstable (bottom) grasps** filtered with an increasing threshold (higher the better). Likelihoods from Grasp Flow (Blue) achieves a more optimal balance between grasp stability and collision.

3.3 Point Cloud Latent Feature Visualization

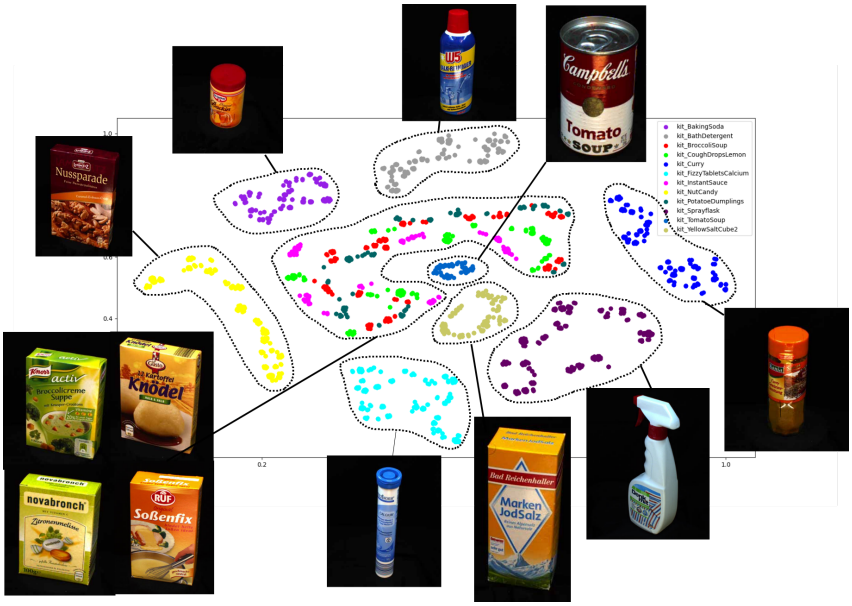
In this subsection, we compare the point cloud latent feature visualization from three models, namely *FFHFlow-cnf*, *FFHFlow-lvm* and FFHNet [2] in Figure 5.

Though *FFHFlow-cnf* has achieved encouraging improvements in terms of diversity and accuracy when compared to the Conditional Variational Autoencoder(cVAE)-based approach, we found *FFHFlow-cnf* less generalizable with limited performance gain. We attribute this problem to the inadequate expressivity of the latent feature, especially when the model needs to understand the complicated relationships between the grasps and the partially observed point clouds of different objects. For example, from our empirical observation, the latent features are assumed to be capable of extracting *two-level hierarchical grasp-relevant information* such as object shape or category from the partially seen object point clouds. (1) *object level* summarizes the grasp-related clues of different objects, such as a box and a bottle; (2) *instance level* subsumes the grasp-associated details of an instance of the same object but captured from different viewpoints.

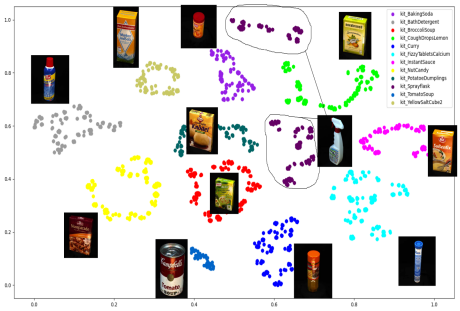
We note that for *FFHFlow-cnf*, we generate the feature visualization with different random seeds to the one in the paper. Nevertheless, both exhibit similar behaviors, further confirming the underperformance of *FFHFlow-cnf* in extracting geometrically meaningful features.

3.4 Experiments of Grasping in Cluttered Scenarios

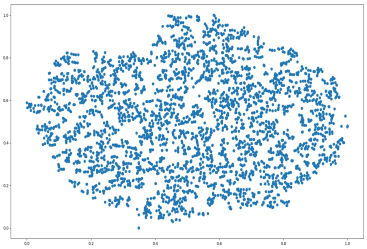
A diverse grasp generator further enables its application for grasping objects in clutter. We propose our grasping pipeline for cluttered unknown objects leveraging Large Language Models (LLMs) and Vision Language Models (VLMs). We first prompt ChatGPT 4o [13] to obtain object names shown on the table and further feed these namse to Grounded SAM [14] to segment out objects. The



(a) *FFHFlow-lvm*



(b) *FFHFlow-cnf*



(c) cVAE

Figure 5: **Point Cloud Feature Visualization** based on t-SNE from (a) *FFHFlow-lvm*, (b) *FFHFlow-cnf* and cVAE in [2]. We illustrate t-SNE features on all 12 KIT test objects used in the simulation. The objects in (a) with similar shapes are closer on the feature space, especially the four boxes on the left bottom part, demonstrating the geometric meaningfulness of the latent features. The cVAE-based approach depicts the least meaningful feature visualization, where the latent samples are drawn from an input-independent prior.

nearest objects will be chosen to be grasped first. We randomly add one obstacle to each different scene to increase the clutterness, to further mimic the household scenarios, show in Figure 6. We convert partial point cloud to meshes and further filter out collisions between environment and the robotic hand using Flexible Collision Library (FCL) [15].

Figure 6: The cluttered scenes contain four unknown objects with additionally unknown obstacles, namely the flower vase and the drawer.

We conducted the grasping experiment for 4 cluttered scenes, each with 4 different objects. We evaluate the cluttered grasping performance with success rate (SR) and clearance rate (CR) in manuscript. The success rate is measured by successful grasps out of all grasp attempts, and the clearance rate is measured by the number of times robots can clear the scene.

FFHFlow-lvm outperforms FFHNet [2] with 7.8% in terms of success rate with a better clearance rate. We observe several failures from FFHNet [2] where a less diverse grasp generator fails to generate valid grasps for occluded objects, especially the blue bowl under the flower vase and the foam brick close to the drill, where the top grasps will be filtered by collision. We further illustrate the influence of diverse grasp distribution in cluttered scenes in Figure 7.

3.5 Visualization of Predicted Grasp Palm Poses and Joints

To show the enhanced diversity, we first compare the grasp palm pose distribution of different approaches shown in Figure 14. By comparing horizontally, we can inspect that our flow-based variational approach, *FFHFlow-lvm* can model the target multi-modal distribution with higher fidelity. Meanwhile, *FFHFlow-cnf* achieves similar results as *FFHFlow-lvm*, especially for box-like objects, but still generates relatively flattened top grasps for cylinder-like objects, such as 2,3,5 rows in 14. In contrast, the cVAE-based approach can only predict less diverse grasps due to the *mode-collapse* problem.

On the other hand, we also visualized the grasps of the full hand, including both the palm poses and hand joint configurations for grasping in clutter in Figure 8 and in Figure 10, single objects in the real-world in Figure 9, from *FFHFlow-lvm*. By inspecting these figures, we can see the dexterity in the predicted hand joints. Moreover, when comparing the grasps from *FFHFlow-lvm* and FFHNet in Figure 8, we can see the diversity of the hand, including the palm and the finger joints, are greater for *FFHFlow-lvm*.

3.6 Failure analysis for Simulation and Real-world Experiments

In the simulation experiment, as shown in Figure 12, *FFHFlow-lvm* causes 2 failures (15.4%) from unstable grasp palm pose, 9 failures (69.2%) from wrong joint configurations, and 2 failures (15.4%) from collisions between the hand and the target object. Failures resulting from joint configurations depict grasps where fingers often are not close enough to apply sufficient force in simulation. This kind of failure normally doesn't exist in real-world experiments. Because the hand impedance controller tends to close the finger more if it's not in contact. However, since this controller cannot be simulated, we replace it with a positional controller. Meanwhile, *FFHFlow-cnf* causes 6 failures from grasp poses, 8 from joint configurations, and 4 from collision. We observe that *FFHFlow-cnf* tends to fail more often because of wrong-predicted grasp poses and collisions. This reason holds for the baseline FFHNet [2] as well (13 failures from grasp poses, 13 from joint configurations, and 7 from collision).

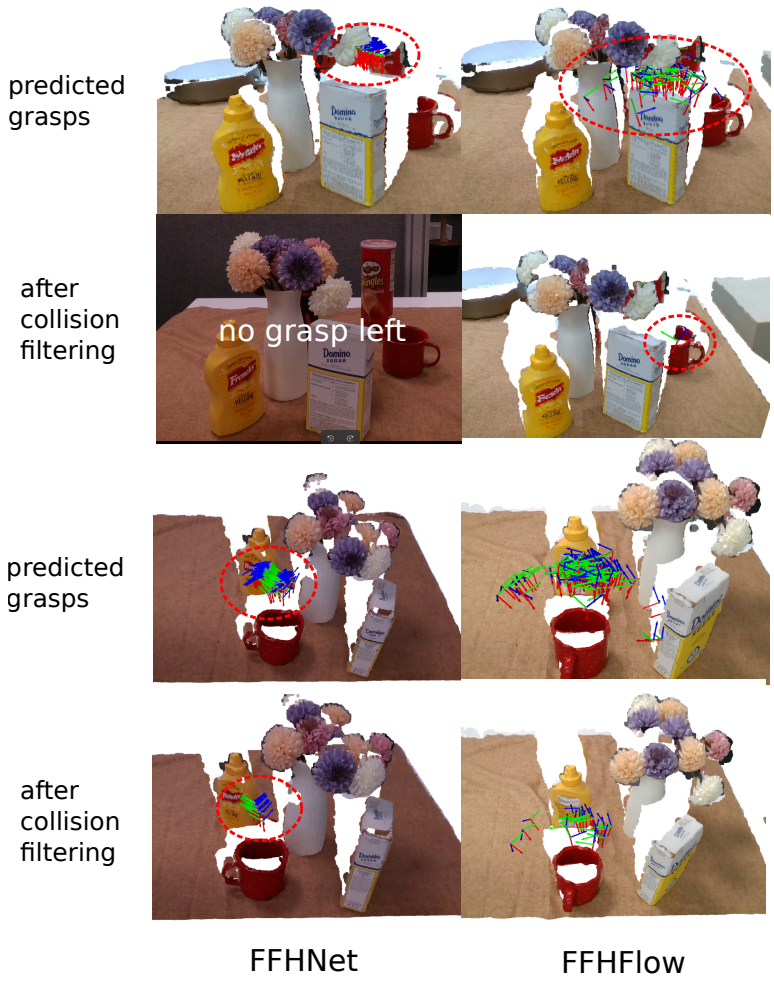


Figure 7: **The grasp distribution generated by FFHNet [2] and *FFHFlow-lvm* before and after collision filtering. A less diverse grasp distribution restricts its application to cluttered scenes.**

In the real-world experiments, by analyzing the errors in Figure 11, when compared to *FFHFlow-cnf* and FFHNet [2], *FFHFlow-lvm* has much fewer failures from unstable grasp pose and a similar number of those from collisions. This trend verifies the superior generalization ability of *FFHFlow-lvm*. On the other hand, for the objects with a low success rate, *FFHFlow-cnf* tends to grasp the corner from a tilted angle instead of the body for the sugar box (40%). FFHNet [2] failed a lot for metal mugs (30%) due to its bias toward top grasps that are harder than side grasps. Apple (30% on average) has the lowest success rate for all models because of its slippery surface, which is often the reason for unstable grasp pose.

3.7 Ablation Study for FFHFlow-cnf

To conduct a fair comparison between *FFHFlow-cnf* and *FFHFlow-lvm* , we increase the size of *FFHFlow-cnf*, namely doubling the layers of the flow. In Table 7, even with a two-times larger size, we can only observe slight improvement, which highlights the inherent limitation on the expressiveness in the latent space of *FFHFlow-cnf*.

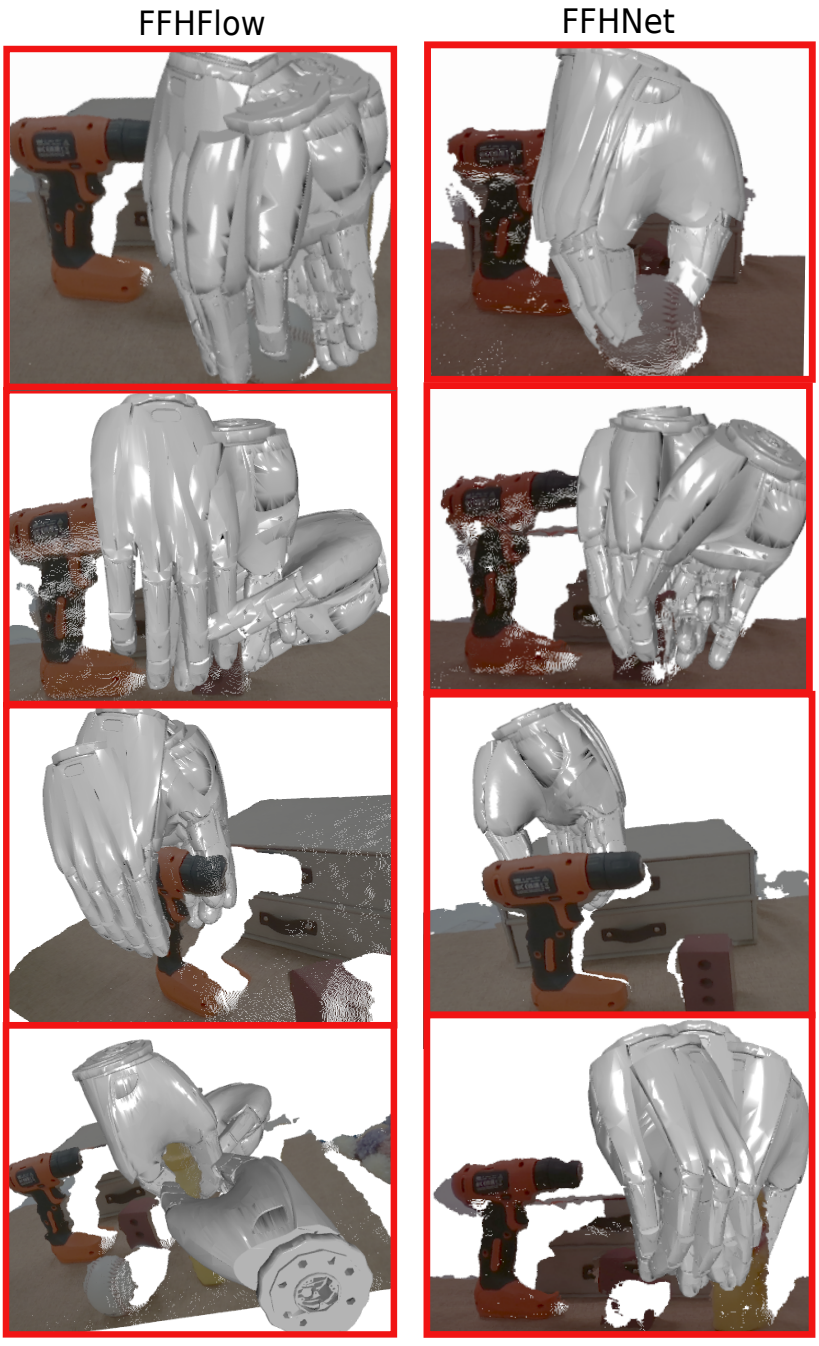


Figure 8: **Comparion on visualization of top 5 scored grasps in the cluttered scene in real-world experiments. FFHFlow demonstrates the ability to generate grasps with better diversity.**

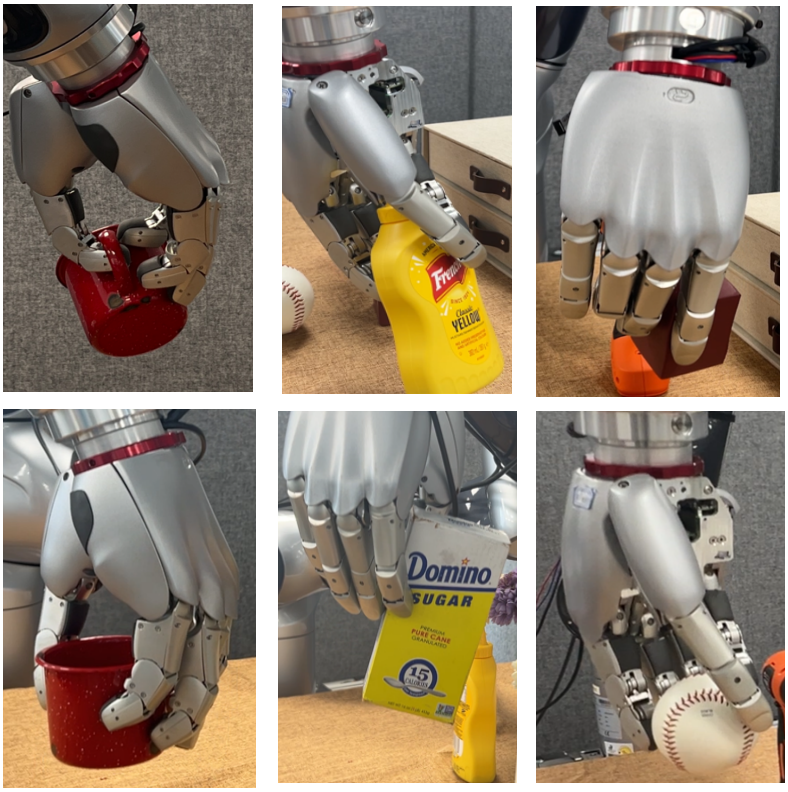


Figure 9: **Exemplar screenshots of grasps in real-world experiments.**

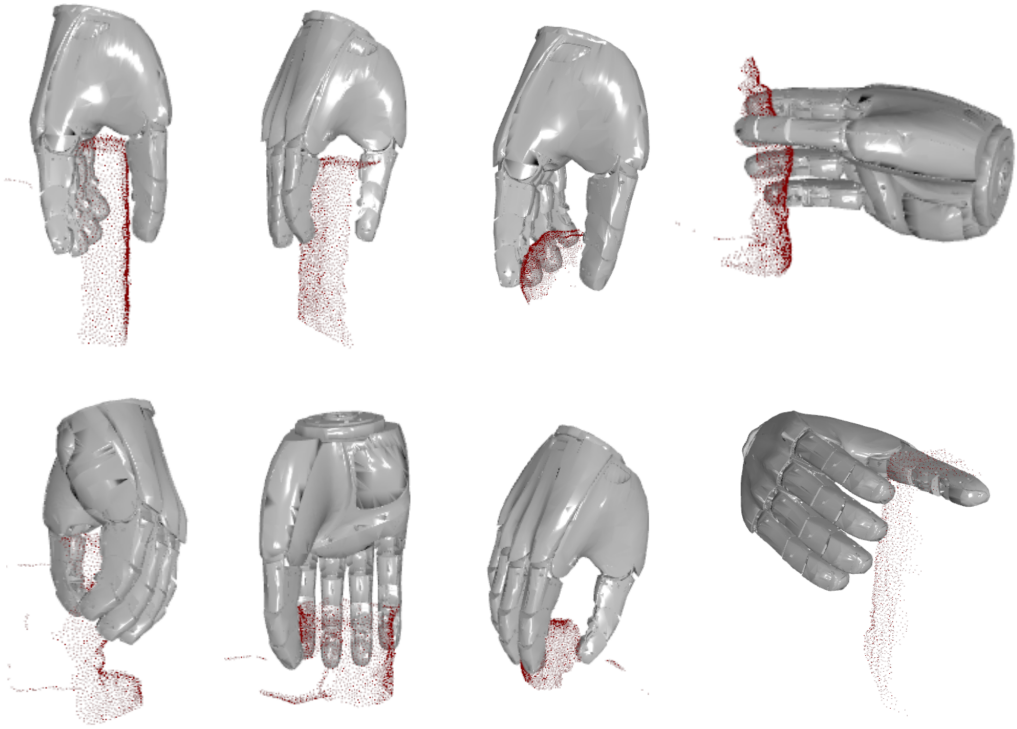


Figure 10: **Exemplar grasp visualization in real-world experiments.**

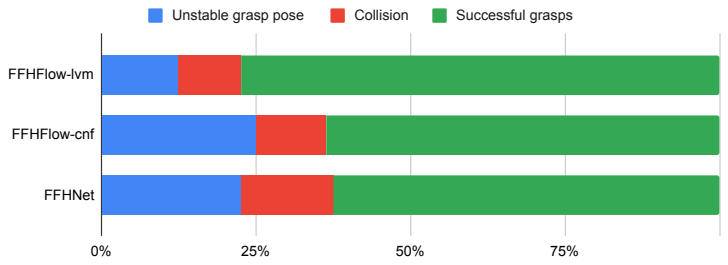


Figure 11: **Failure Analysis for the Real-world Experiment**

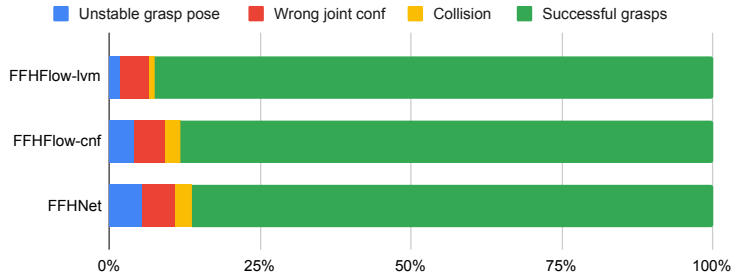


Figure 12: **Failure Analysis for the Simulation Experiment**

Table 7: Ablation Study for FFHFlow-cnf

Methods (size)	Objects												Average Succ Rate
	Baking Soda	Bath Detergent	Broccoli Soup	Cough Drops	Curry	Fizzy Tablets	Instant Sauce	Nut Candy	Potato Dumpling	Spray Flask	Tomato Soup	Yellow SaltCube	
FFHFlow-cnf (8 layers)	95.0%	100.0%	95.0%	90.0%	100.0%	80.0%	95.0%	85.0%	100.0%	100.0%	70.0%	70.0%	92.5%
FFHFlow-cnf (16 layers)	95.0%	90.0%	95.0%	95.0%	100.0%	95.0%	90.0%	90.0%	95.0%	95.0%	90.0%	85.0%	92.9%

3.8 Ablation Study of FFHFlow-lvm

The question of "What are the critical factors in the proposed models that influence the performance most?" is of particular interest for better understanding the proposed models, in particular *FFHFlow-lvm* for its complexity. From the results of an ablation study in Table 8, we can draw the messages that positional encoding pre-processing, adding conditional base distribution only to grasp flow generator can help alleviate over-fitting and improve the generalization performance. Here positional encoding is applied on Euler angles (3D) to obtain 60D, compared to the baseline of using 6D rotation representation [16], originally already used in FFHNet [2].

The potential reason for the benefit of positional encoding can be the better capability of expressing high-frequency information from low-dimensional data such as 3-d angel vectors in our case [17]. Moreover, the size ratio of two flows in the model seems to influence the training stability. When the model has two different number of layers assigned to the grasp and prior flow, the coverage is much lower and the simulation evaluation failed due to some feasible predicted values from the model.

Evaluating Predicted Finger Joints To investigate how much the predicted finger joints matter, we conduct an ablation study on comparing the success rate of grasping with and without the predicted joints in simulation. In case of grasping without predicted joints, we set the corresponding joints with $0.2\ rad$ to approximate a power grasp for each object. In Table 9, we can observe a clear drop when grasping without the predicted joints for both methods, confirming their positive effects for precise grasp synthesis. Moreover, the increase brought by *FFHFlow-lvm* (9.6%) is significantly higher than that of the cVAE approach [2] (1.6%). Such improvement can demonstrate the overall benefits of our proposed models for not only the predicted palm poses but also the predicted joints.

Table 8: Ablation study of *FFHFlow-lvm* on Cov and Success rate

Ablated Models (w/o eval)	Cov ↑	Succ Rate ↑
FFHFlow-lvm	30.2%	94.6%
6D (w/o-positional-encoding)	30.1%	92.3%
both-flows-cond-base	30.4%	89.8%
both-flows-w/o-cond-base	30.7%	88.2%
grasp-flow-4-layers	28.8%	-
both-flows-4-layers	30.2%	93.3%

Table 9: Predicted Joints Evaluation

Methods	Objects												Average Succ Rate
	Baking Soda	Bath Detergent	Broccoli Soup	Cough Drops	Curry	Fizzy Tablets	Instant Sauce	Nut Candy	Potato Dumpling	Spray Flask	Tomato Soup	Yellow SaltCube	
FFHNet [2] w/o Joints	95.0%	95.0%	95.0%	90.0%	95.0%	100.0%	75.0%	60.0%	80.0%	95.0%	65.0%	70.0%	84.6%
FFHNet [2]	90.0%	95.0%	100.0%	95.0%	95.0%	95.0%	100.0%	65.0%	85.0%	90.0%	65.0%	60.0%	86.2%
FFHFlow-lvm w/o Joints	90.0%	70.0%	100.0%	90%	100.0%	90.0%	85.0%	85.0%	85.0%	75.0%	65.0%	85.0%	85.0%
FFHFlow-lvm	95.0%	95.0%	95.0%	100.0%	95.0%	95.0%	100.0%	85.0%	100.0%	90.0%	100.0%	85.0%	94.6%

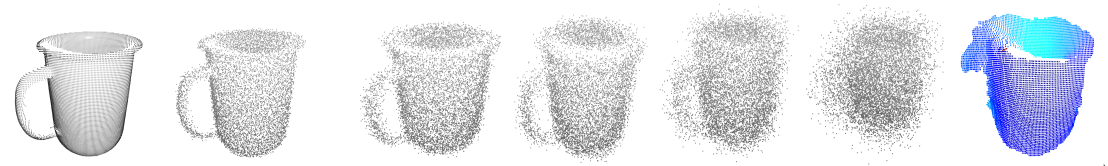
3.9 Influence of Point Cloud Noises to FFHFlow-lvm

We add random gaussian noise to the point cloud in simulation and feed it to *FFHFlow-lvm*. The results in Table 10 demonstrate its negative influence of noises on success rate. We observed a roughly linear performance drop between 0mm to 5mm and then the performance drops dramatically from 5mm with 75.3% to 10mm with 29.9%. Given a real world point cloud in Table 10, it's noise level is estimated to be between 0 and 1 mm. Therefore we can expect a performance drop of around $3-4\%$ given same level of noise from real world, ignoring all other sim2real gap. Other sim2real gap for point cloud could be missing pixels from physical camera and imperfect segmentation mask.

To better minimize the negative influence of noise or improve the robustness against noise, we could in principle train the model with the simulated noised point cloud.

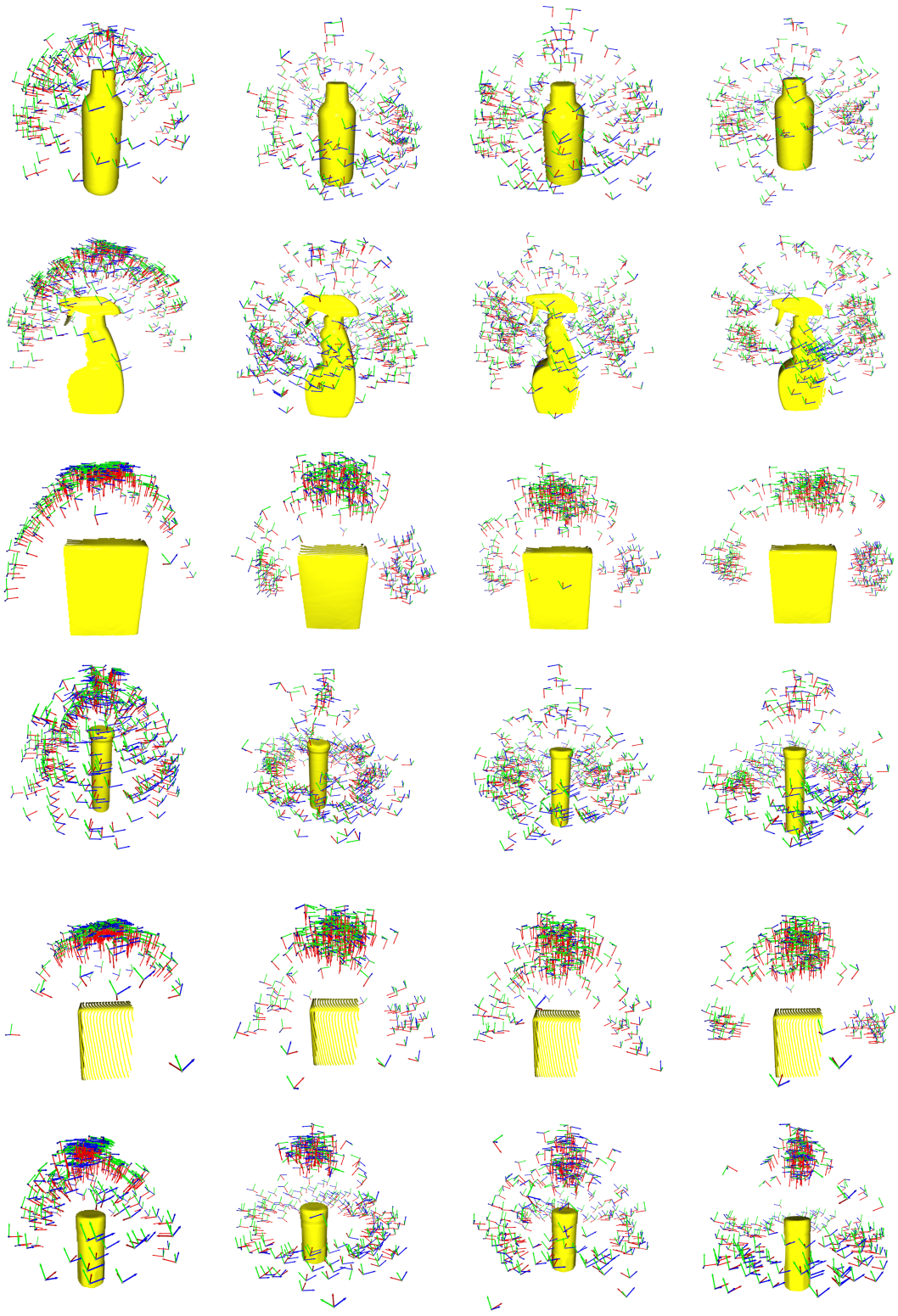
Table 10: Success Rate Drop vs Point Cloud Noise

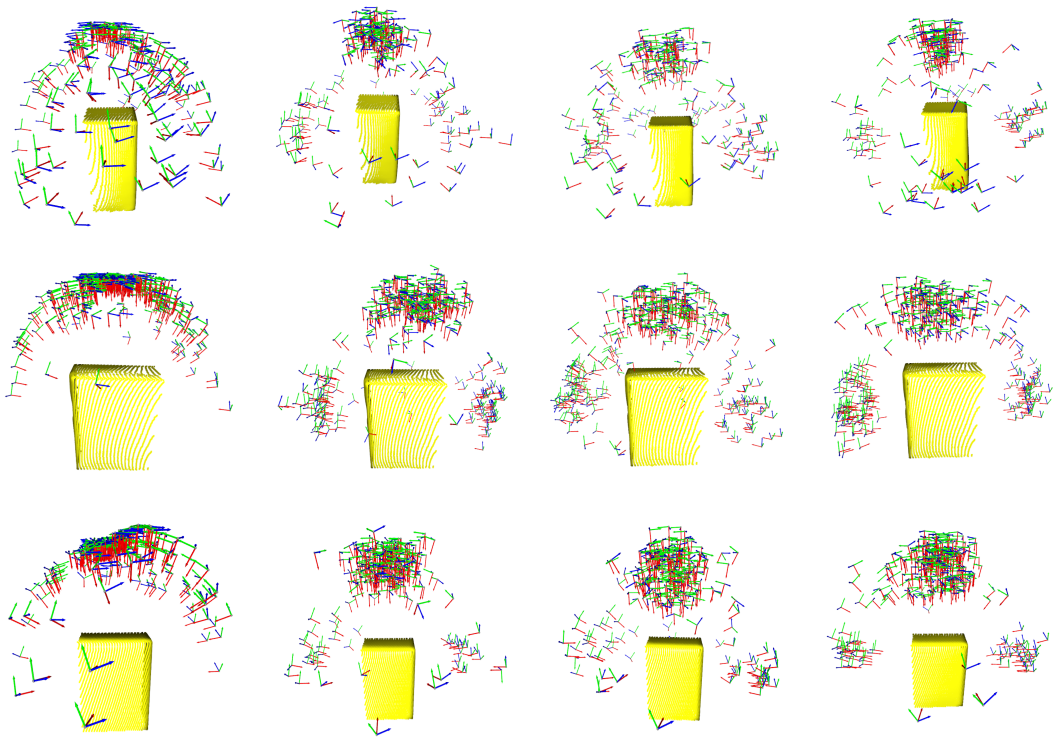
	Gaussian Noise					
Standard deviation	0 mm	1mm	2mm	3mm	5mm	10mm
FFHFlow-lvm	94.6%	91.2%	88.3%	83.3%	75.3%	29.9%



(a) Original point cloud (b) 1mm noise (c) 2mm noise (d) 3mm noise (e) 5mm noise (f) 10mm noise (g) Real world point cloud

Figure 13: **Visualization of Point Cloud applied with different magnitude of Gaussian Noise.** We apply noise generated from a zero mean and a parameterized standard deviation Gaussian distribution to original point cloud. The added standard deviation is in a range from 1mm till 10mm. We can see the point cloud gets more fuzzy and almost not recognizable after 5mm. Compared to (g) real world point cloud, we can estimate its noise level is between 0-1 mm standard deviation.





(a) cVAE (b) *FFHFlow-cnf* (c) *FFHFlow-lvm* (d) Ground truth

Figure 14: **Visualization of grasp pose distributions** from (a) cVAE in [2], (b) *FFHFlow-cnf*, and (c) *FFHFlow-lvm* and (d) ground truth.

References

[1] A. Kasper, Z. Xue, and R. Dillmann. The kit object models database: An object model database for object recognition, localization and manipulation in service robotics. *The International Journal of Robotics Research*, 31(8):927–934, 2012.

[2] V. Mayer, Q. Feng, J. Deng, Y. Shi, Z. Chen, and A. Knoll. Ffhnet: Generating multi-fingered robotic grasps for unknown objects in real-time. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 762–769, 2022. doi:10.1109/ICRA46639.2022.9811666.

[3] M. T. Ciocarlie and P. K. Allen. Hand posture subspaces for dexterous robotic grasping. *The International Journal of Robotics Research*, 28(7):851–867, 2009.

[4] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981.

[5] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 international conference on advanced robotics (ICAR)*, pages 510–517. IEEE, 2015.

[6] S. Prokudin, C. Lassner, and J. Romero. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4332–4341, 2019.

[7] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017.

[8] V. Stimper, D. Liu, A. Campbell, V. Berenz, L. Ryll, B. Schölkopf, and J. M. Hernández-Lobato. normflows: A pytorch package for normalizing flows. *arXiv preprint arXiv:2302.12014*, 2023.

[9] D. P. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

[10] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[11] Q. Feng, D. S. M. Lema, M. Malmir, H. Li, J. Feng, Z. Chen, and A. Knoll. Dexgangrasp: Dexterous generative adversarial grasping synthesis for task-oriented manipulation, 2024.

[12] Z. Weng, H. Lu, D. Kragic, and J. Lundell. Dexdiffuser: Generating dexterous grasps with diffusion models, 2024.

[13] OpenAI. Chatgpt (july 11 version), 2024. URL `https://www.openai.com`. Large language model.

[14] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, and L. Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024.

[15] J. Pan, S. Chitta, and D. Manocha. Fcl: A general purpose library for collision and proximity queries. In *2012 IEEE International Conference on Robotics and Automation*, pages 3859–3866, 2012. doi:10.1109/ICRA.2012.6225337.

[16] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li. On the continuity of rotation representations in neural networks, 2020. URL `https://arxiv.org/abs/1812.07035`.

[17] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020.