Decomposed Vector-Quantized Variational Autoencoder for Human Grasp Generation

Zhe Zhao[1,2,3], Mengshi Qi[1,2,3*], and Huadong Ma[1,2,3]

[1] State Key Laboratory of Networking and Switching Technology
[2] Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia
[3] Beijing University of Posts and Telecommunications, China

Abstract. Generating realistic human grasps is a crucial yet challenging task for applications involving object manipulation in computer graphics and robotics. Existing methods often struggle with generating fine-grained realistic human grasps that ensure all fingers effectively interact with objects, as they focus on encoding hand with the whole representation and then estimating both hand posture and position in a single step. In this paper, we propose a novel Decomposed Vector-Quantized Variational Autoencoder (DVQ-VAE) to address this limitation by decomposing hand into several distinct parts and encoding them separately. This part-aware decomposed architecture facilitates more precise management of the interaction between each component of hand and object, enhancing the overall reality of generated human grasps. Furthermore, we design a newly dual-stage decoding strategy, by first determining the type of grasping under skeletal physical constraints, and then identifying the location of the grasp, which can greatly improve the verisimilitude as well as adaptability of the model to unseen hand-object interaction. In experiments, our model achieved about 14.1% relative improvement in the quality index compared to the state-of-the-art methods in four widely-adopted benchmarks. Our source code is available at `https://github.com/florasion/D-VQVAE`.

Keywords: Grasp Generation · Decomposed Architecture · Variational Autoencoder

1 Introduction

Generating how a human would grasp different given objects [12–14, 40] has already become crucial in many domains, including robotics, human-computer interaction, and augmented reality applications. Although 3D hand pose estimation [1,4,7,41] and hand-object 3D reconstruction [5,11,34] have made significant strides in recent years, grasp generation extends beyond them and becomes imperative to have a finer-grained comprehension of the individual components of the highly articulated human hand.

* Corresponding author: qms@bupt.edu.cn.

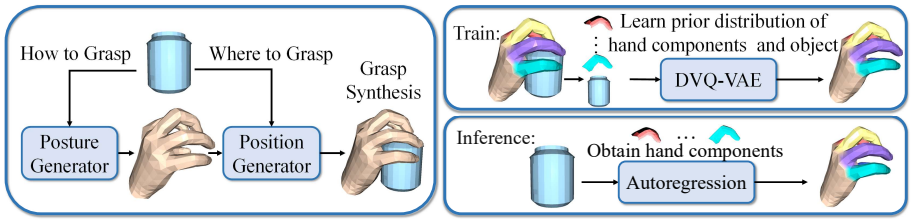

Fig. 1: Illustration of the proposed grasp generation model. Initially, we utilize Decomposed VQ-VAE (DVQ-VAE) to learn the prior distributions of the object and each hand component (*i.e.*, five fingers and the palm) during training. Specifically, we divide the decoding process into two stages hand posture and position generation. During inference, we perform autoregression using the object as a guide to obtain the realistically generated human grasp.

To generate high-quality grasps, the main challenge lies in accurately modeling the interaction between the hand and objects. In recent years, numerous efforts [12–14,40] have been adopted into this field, which is primarily focused on predicting the contact map to ensure the hand comes into full contact with the object. However, such methods often lead to generated unreasonable grasps that resemble touches rather than grasps [16]. In addition, existing methods predominantly employ Conditional Variational Autoencoders (CVAE) [12–14,16,32,40], while some have incorporated Generative Adversarial Networks (GANs) [3, 8]. But they utilize a continuous latent space to model all kinds of different grasps, which does not reflect the discrete and categorical nature of real-life human hands, resulting in poor diversity and fallacious grasps. In contrast, Vector Quantized Variational Autoencoder (VQ-VAE) [35] can lead to a more controllable generation of grasps by directly utilizing discrete latent codes instead of randomly sampling from a Gaussian distribution. Therefore, we opt for VQ-VAE [35] as our core network and extend it in a new decomposed manner to take full advantage of each hand component and derive more diversity from the autoregressive inference.

Another challenge is that existing methods cannot fit well into unseen or out-of-domain objects. Several research proposes test-time adaptation [12, 13], which is an optimization-based refinement during inference by directly optimizing MANO parameters [30]. However, such approaches will lead to a significant time cost. To mitigate the time overhead during testing, we can break down the complex process of grasping into two distinct stages, *i.e.*, predicting the hand pose by considering the possible grasp type and then finding the suitable position for the manipulated object.

In this paper, we present a novel Decomposed VQ-VAE (DVQ-VAE) to generate human grasps that exhibit both diversity and authenticity when given an object. Our innovation lies in our model can partition the hand into multiple components for encoding them into several discrete latent spaces, and then executing grasp generation by the newly designed dual-stage decoding strategy.

Specifically, we propose a part-aware decomposed architecture to decouple the hand into six distinct components including five fingers and the palm, each of which is encoded into its respective codebook. Furthermore, grasp generation occurs in two sequential steps by initially generating the posture of the hand followed by the determination of the grasp location.

Our main contributions can be summarized as follows:

(1) We propose a novel DVQ-VAE for human grasp generation, which is based on a part-aware decomposed architecture to encode multi-part of the hand and allows for progressively generating the complete hands. To the best of our knowledge, we are the first to propose such VQ-VAE [35] based framework for this specific task.

(2) We introduce a new dual-stage decoding strategy to improve the quality of generated grasps for unseen objects, by gradually determining the grasp type under skeletal physical constraints and its position.

(3) We benchmark our proposed method on the four popular datasets, with 18% relative improvements over the state-of-the-art approaches in the penetration volume. More importantly, we achieve nearly 14.1% relative improvement in the quality index metric.

2 Related Work

Grasp Generation. With the development of virtual reality and robotics technology, significant efforts [3, 12, 13, 16–18, 25, 26] have been devoted to generating the physically plausible and diverse grasps, most of which are based on the generative models, including GAN [3, 8, 28, 37] and VAE [12, 13, 15, 16, 40]. A representative work is Jiang *et al.* [12] introduced a CVAE [31] to generate grasps, taking objects into account as conditions. This approach incorporated test-time adaptation by using ContactNet [12]. Furthermore, VQ-VAE [35] has demonstrated its superiority in multiple areas, such as image generation and 3D synthesis [20, 23]. However, it is noteworthy that, to the best of our knowledge, there exists no prior research that has applied VQ-VAE [35] to the field of grasp generation. Drawing inspiration from the work of Pi *et al.* [23], who segmented the human body into five parts and encoded them into multiple codebooks for generating human motions, we extended this concept to encode the hand into multiple components. Through the utilization of VQ-VAE [35], we encoded the data as discrete variables, thereby enabling us to acquire a prior distribution of hand-object interactions and then exploit it in the grasp prediction.

Hand-Object Interaction. Modeling hand-object interaction [2, 9, 11, 27, 30, 33, 38, 39] is a crucial task, but the considerable degrees of freedom in hand pose present significant challenges. A pioneer work is Romero *et al.* [30] proposed the parameterized hand model MANO to address the difficulties in reconstructing hand pose and shape. Different from modeling hand-object interaction assumes that both the hand and the object are visible in the input, we aim to predict a realistic grasp for the observed objects without seeing the hand.

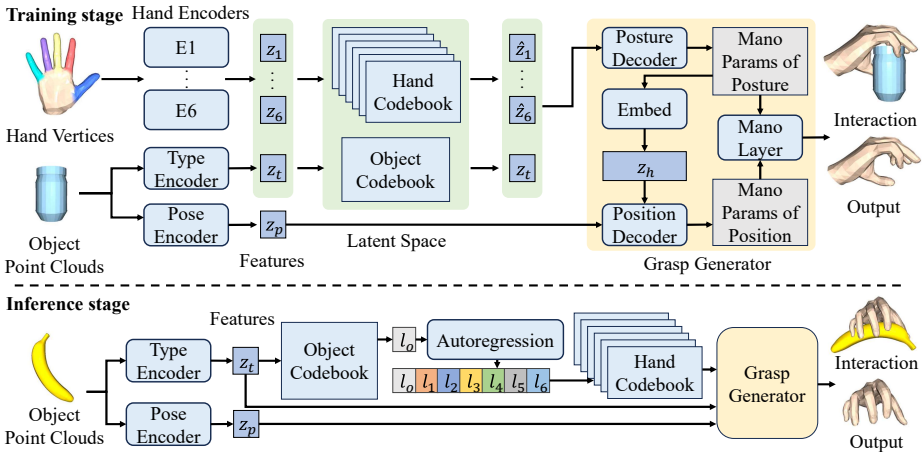

Fig. 2: Overall architecture of the proposed DVQ-VAE model which is based on the encoder-decoder paradigm. During training, the model takes both hand vertices and object point clouds as inputs and maps them into discrete latent spaces consisting of seven codebooks (*i.e.*, one for object and six for hand) based on different hand components to generate hands. While, at the inference phase, we only use object point clouds as input to generate hands capable of grasping the given object.

3 Proposed Approach

In our work, our goal is to generate a stable and physically plausible hand mesh for grasping by giving an object point cloud as input. To achieve this, we design a novel Decomposed VQ-VAE and divide the decoding process into two stages to obtain the final grasps, as illustrated in Fig. 2.

3.1 Overview

Training. As shown in the top part of Fig. 2, given the sampled object points $P^o \in \mathbb{R}^{N_o*3}$ and hand vertices $P^h \in \mathbb{R}^{778*3}$ as input, we utilize two encoders to extract object features, denoted as z_t and z_p. Simultaneously, we partition the hand vertices into $N(N = 6)$ segments, comprising five distinct finger components (*i.e.*, thumb, index finger, middle finger, ring finger, and little finger) and the palm, resulting in six distinct segments that are individually encoded to obtain hand features $z_f = \{z_1, ..., z_N\}$. These hand features, along with z_t, are projected into an discrete latent embedding space where $N + 1$ separate codebooks $Z = \{Z_o, Z_1, ..., Z_N\}$ are learned. During decoding, We use MANO [30] to generate hand mesh, and MANO [30] has several sets of parameters, *i.e.*, $M_\alpha \in \mathbb{R}^{10}$ handles person-specific hand shape, $M_\beta \in \mathbb{R}^{45}$ manages joint rotations, $M_\gamma \in \mathbb{R}^3$ controls hand translation, and $M_\delta \in \mathbb{R}^3$ deals with hand rotation. Among these, M_α and M_β are related to hand posture, and we combine them denoted as posture $\hat{M}_{posture}$. While M_γ and M_δ are related to hand position, and

we combine them referred to position $\hat{M}_{position}$. The returned closest matching vectors $\hat{z}_f = \{\hat{z}_1, ..., \hat{z}_N\}$ of the hand features from Z are then concatenated with z_t and fed into the Posture Decoder to decode $\hat{M}_{posture}$. Then, we concatenate the encoded result of $\hat{M}_{posture}$, *i.e.*, z_h, with z_p and directed into the Position Decoder to obtain $\hat{M}_{position}$. These two sets of MANO [30] parameters are then processed through the MANO [30] layer to derive the final hand mesh.

Inference. As shown in the bottom part of Fig. 2, our model utilizes the sampled object points $P^o \in \mathbb{R}^{N_o * 3}$ as input, and achieves object features, z_t and z_p through the two object encoders. Then, z_t is matched with the nearest vector in the object codebook Z_o to obtain an index l_o. Subsequently, we adopt an autoregressive model conditioned on l_o to generate a sequence of hand codebook indices $l_h = \{l_1, ..., l_N\}$. These hand codebook indices are then matched in the learned hand codebooks to obtain the corresponding hand features $\hat{z}_f = \{\hat{z}_1, ..., \hat{z}_N\}$. Finally, these hand features, along with z_t and z_p, are fed into the grasp generator to produce the desired hand mesh.

In our model, PointNet [24] serves as the point cloud encoder for both objects and hands, and all decoders are MLPs. For the inference phase, we employ PixelCNN [22] as the self-regressive model.

3.2 Object Encoder

Different from existing methods [12–14, 16, 32, 40], we propose two object encoders: a type encoder and a pose encoder. The idea behind the type encoder is that all grasp types are uncountable and each of them is relevant to the shape of the given object, we can extract the z_t to help learn an object codebook during the latent space embedding. And the codebook records these types and even clusters them into several common grasp categories so that our model can know how to grasp any input object. While, the pose encoder focuses on extracting features that aid in decoding grasp positions, which will determine where the hand should contact with the object.

3.3 Part-Aware Decomposed Architecture

Here we describe the proposed part-aware decomposed architecture of DVQ-VAE. Traditional VQ-VAEs [35] are typically used for image generation and employ a single codebook. This stems from the same codebook indices can be located in various positions of the image, while during autoregression the same codebook also can be used to infer the pixel at any position. However, in the case of human hands, the positions of different fingers are fixed. Therefore, we propose to extend VQ-VAE [35] to a part-aware decomposed architecture, *i.e.*, encoding the object and the N parts of the hand as z_t and z_f, respectively, and search for the vectors with the minimum Euclidean distance in their respective discrete codebooks during the latent embedding. To prevent the loss of object features, we use the discovered $\hat{z}_f$ in combination with the original z_t as input for the grasp generator. In this scenario, the object's latent variable serves solely

as a condition during autoregression. Thus, the training of the object codebook is conducted through unsupervised learning. In this context, to encourage z_t and z_f to closely approximate the matched vectors in the codebooks, the loss function $\mathcal{L}_E$ for codebook training can be defined as follows:

$$\mathcal{L}_h = \sum_{i=1}^{N} \|sg(\hat{z}_i) - z_i\|_2^2 + \beta \sum_{i=1}^{N} \|sg(z_i) - \hat{z}_i\|_2^2, \tag{1}$$

$$\mathcal{L}_o = \|sg(\hat{z}_t) - z_t\|_2^2 + \beta \|sg(z_t) - \hat{z}_t\|_2^2, \tag{2}$$

$$\mathcal{L}_E = \lambda_e \cdot (\mathcal{L}_h + \mathcal{L}_o), \tag{3}$$

where

$$\hat{z}_i = e_k, \text{ where } k = argmin_j \|z_i - e_j\|_2, \tag{4}$$

$$\hat{z}_t = e_m, \text{ where } m = argmin_j \|z_t - e_j\|_2, \tag{5}$$

where the operator $sg(\cdot)$ signifies cessation of the gradient flow, impeding the propagation of gradients into its associated parameter, β is the hyperparameter set to 0.25, and λ_e is the hyper-parameter, e represents the embedding in the codebook.

During inference, we follow VQ-VAE [35] to employ PixelCNN [22] as our autoregressive model, and we use the object's codebook indices as both the condition and the initial sequence for PixelCNN [22] to predict a sequence of codebook indices for the hand components.

3.4 Dual-Stage Decoding Strategy

Previous methods [12, 16] improved grasp quality by generating 61 MANO [30] parameters in one step and optimizing these parameters over multiple steps, which resulted in significant time consumption. In order to overcome the weaknesses of single-step decoding, we design a dual-stage decoding strategy by generating the grasp posture and grasp position in sequence. In the spirit of decoupling, we divide the MANO parameters into two parts: position and posture, and we first generate the more numerous grasp posture parameters $\hat{M}_{posture}$ and then the fewer position parameters $\hat{M}_{position}$ by combining hand features with the object's characteristics.

1) Firstly, to generate a reasonable grasp during decoding, we utilize the L2 distance of the ground truth $M_{posture}$ and the predictions:

$$\mathcal{L}_{posture} = \left\| M_{posture} - \hat{M}_{posture} \right\|_2. \tag{6}$$

Furthermore, we introduce physical constraints based on hand-skeletal dynamics. We extract skeletal key points J from the reconstructed hand and calculate

angles between adjacent key points. From these angles, we generate gating information to correct grasps that do not conform to hand-skeletal dynamics:

$$\theta_i = \cos^{-1}\left(\frac{\overrightarrow{J_iJ_{i-1}} \cdot \overrightarrow{J_iJ_{i+1}}}{\|\overrightarrow{J_iJ_{i-1}}\|\|\overrightarrow{J_iJ_{i+1}}\|}\right), \tag{7}$$

$$\hat{M}_{posture} = \hat{M}_{posture} + G(\theta) \odot T(\hat{M}_{posture}), \tag{8}$$

where θ represents the joint angles, $\overrightarrow{J_iJ_{i+1}}$ represents the vector from key point J_i to J_{i+1}, $G(\cdot)$ denotes the network used to generate gating information, and $T(\cdot)$ stands for the transformer layer used to produce correction values.

2) Secondly, for training the position decoder, we stop the gradient propagation at z_h and compute the L2 distance of $M_{position}$ and hand vertices formulated as:

$$\hat{M}_{position} = Dec[sg(z_h), z_p], \tag{9}$$

$$\mathcal{L}_{position} = \left\| M_{position} - \hat{M}_{position} \right\|_2, \tag{10}$$

$$\mathcal{L}_v = \left\| P^h - M(\hat{M}_{position}, \hat{M}_{posture}) \right\|_2, \tag{11}$$

where the operator $Dec[\cdot]$ represents decoding through the position decoder, and the operator $M(\cdot)$ represents passing through the MANO layer [30]. Then the reconstruction loss L_R is obtained by combining $\mathcal{L}_{posture}$, $\mathcal{L}_{position}$, and $\mathcal{L}_v$:

$$\mathcal{L}_R = \lambda_h \cdot (\mathcal{L}_{posture} + \mathcal{L}_{position}) + \lambda_v \cdot \mathcal{L}_v, \tag{12}$$

where λ are hyper-parameters. To prevent the model from inadvertently learning hand position information before the position decoder, we center the hand vertices by subtracting their mean coordinates during training.

3.5 Optimization

Finally, our total objective loss consists of three parts, *i.e.*, $\mathcal{L}_E$ for the discrete latent embeddings in Sec. 3.3, $\mathcal{L}_R$ for constraining the morphology of generated hand in Sec. 3.4, and $\mathcal{L}_{contact}$ for constraining contact between the hand and the manipulated object that we will describe below.

Following [12], we use object-centric contact loss $\mathcal{L}_c$ and contact map consistency loss $\mathcal{L}_m$ to enhance the contact between the hand and the object formulated as:

$$\mathcal{L}_c = \sum_{p_m \in P_m} \min_{p_c \in P_c} |p_m - p_c|, \tag{13}$$

$$\mathcal{L}_m = \frac{|P_m| \cap \left|\hat{P}_m\right|}{|P_m|}, \tag{14}$$

where the operator $|\cdot|$ indicates the number of points in the point set. P_m and $\hat{P}_m$ represent the point set of the ground truth and the predicted grasp contact map, respectively. P_c denotes the points on the hand that could potentially contact the object. p_c and p_m refer to each point in the point set P_c and P_m, respectively.

We also employ penetration loss to enhance the physical reality of the generated grasp, formulated as:

$$\mathcal{L}_p = \sum_{p \in P_{in}} \|p - P_i^o\|_2^2, \tag{15}$$

where

$$P_i^o = \text{argmin}_{p_i \in P^o} \|p - p_i\|, \tag{16}$$

where P_{in} is the subset of hand points that enter into the object. p and p_i denote as each point in the point set P_{in} and P^o, respectively. Hence, we can obtain the contact loss as follows:

$$\mathcal{L}_{contact} = \lambda_m \cdot \mathcal{L}_m + \lambda_c \cdot \mathcal{L}_c + \lambda_p \cdot \mathcal{L}_p. \tag{17}$$

Where λ are hyper-parameters.

Finally, we combine $\mathcal{L}_{contact}$ with $\mathcal{L}_E$ and $\mathcal{L}_R$ to form our overall loss function:

$$\mathcal{L} = \mathcal{L}_R + \mathcal{L}_E + \mathcal{L}_{contact}. \tag{18}$$

4 Experiments

4.1 Datasets

Obman Dataset [11] constitutes a comprehensive compilation of manipulation interaction synthesized data, produced by GraspIt [19]. It encompasses a vast collection of 150,000 grasps across 2,772 distinct objects. Following [12, 16], 141,550 of these grasps are designated for training purposes, while the remaining 6,285 are allocated for testing.

HO-3D [10], FPHA [6] and GRAB [32] dataset. Following [12], due to the Obman dataset [11] containing the greatest variety of objects, we chose to train on the Obman dataset [11] but test on these three datasets. The test data can be regarded as out-of-domain objects used to evaluate the adaptability. Specifically, HO3D [10] contains 10 objects, FPHA [6] contains 4 objects, and GRAB [32] contains 51 objects from 10 subjects.

4.2 Metrics

For a fair comparison, following [11–14, 16, 33], we evaluate the results in terms of the below metrics, of which the *quality index* is our proposed new metric.

1) Contact Ratio (%). It calculates the ratio of grasps that can contact the given object in all generated grasps.

2) Hand-Object Interpenetration Volume (cm^3). It is the volume shared by the 3D voxelized models of the hand and object. Following [13,16], We voxelize the mesh of both hand and object with size 0.1 cm^3.

3) Grasp Simulation Displacement (Grasp Disp) (cm). It quantifies the stability of the generated grasp when the simulated gravity is added, and we report the average displacement of the object's center of mass.
4) Entropy and Cluster Size. Following [13, 16], we divide the generated grasps into 20 clusters using K-means and then measured the entropy and averaged cluster size to assess the diversity of generated grasps. Higher entropy values and larger cluster sizes indicate more diversity.
5) Time (s). It evaluates the inference speed of the model in generating one batch of grasps.
6) Quality Index. Note that we observe the visualized results with only low penetration or low physical simulation displacement does not necessarily indicate it is a high-quality grasp. For example, insufficient contact can reduce penetration but also decrease the stability of the grasp, while the severe penetrating of the hand into the object can reduce displacement in the physical simulator. Inspired by [36], to more scientifically quantify the quality of grasping, we adopt a utility function:

$$Q = a \cdot x + (1 - a) \cdot y, \tag{19}$$

where a represents the weight, used to balance the gap between grasp displacement and penetration,x represents penetration volume, y represents grasp disp. According to [36], we measure the displacement and penetration for each grasp in the obman and grab datasets. And we calculate $a = 0.301$ using the same method as described in [36].

4.3 Compared Methods

We compare our proposed DVQ-VAE against the following state-of-the-art methods: 1) **GraspTTA [12]**: Trained on the Obman dataset [11] and utilizes CVAE [31] to generate grasps conditioned on the object, and further refines the generated grasps using ContactNet with test-time adaptation (TTA). 2) **GraspCVAE [12]**: A variant of GraspTTA without TTA. 3) **ContactGen [16]**: Trained on the Grab dataset [32] and utilizes CVAE [31] to generate contact maps of the object, and then optimizes the hand parameters based on them. 4) **Grasping Field [14]**: Trained on the Obman dataset [11] and utilizes VAE [15] to generate grasps based on the signed distance fields of the object and the hand.

4.4 Implementation Details

We train our model on Obman dataset [11], and sample $N_o = 3000$ points from the object mesh as input. We employ Adam optimizer with an initial learning rate of 1e-4 for 200 epochs, halving the learning rate at epochs 60, 120, 160, and 180, and set $\lambda_e = 10, \lambda_m = -50, \lambda_c = 1500, \lambda_p = 5, \lambda_h = 0.1, \lambda_v = 10$. For training PixelCNN [22], we use Adam optimizer with a learning rate of 3e-4 for 100 epochs. We implement our model based on Pytorch on a single NVIDIA RTX 3090 GPU, the training time is 1000 minutes.

Dataset	Method	Contact Ratio (%)↑	Penetration Volume (cm^3)↓	Grasp Disp (cm) ↓	Time (s) ↓	Entropy ↑	Cluster Size ↑	Quality Index↓
HO-3D [10]	GraspCVAE [12]	99.60	7.23	2.78	**0.0040**	**2.96**	0.81	4.12
	GraspTTA [12]	**100**	9.00	**2.65**	19.67	2.87	0.80	4.56
	ContactGen [16]	90.10	6.53	3.72	119.4	2.94	**4.79**	4.57
	GraspingField [14]	89.60	20.05	4.14	57.49	2.91	3.31	8.93
	Our DVQ-VAE	99.50	**5.36**	2.75	0.14	2.80	3.84	**3.54**
FPHA [6]	GraspCVAE [12]	98.98	7.46	2.97	**0.0038**	2.91	0.81	4.32
	GraspTTA [12]	**100**	8.26	**2.75**	19.29	**2.93**	0.76	4.41
	ContactGen [16]	94.00	10.43	3.64	237.38	2.84	2.88	5.68
	GraspingField [14]	97.00	29.78	5.47	58.57	2.85	2.36	12.79
	Our DVQ-VAE	97.96	**4.58**	3.35	0.14	2.86	**3.53**	**3.72**
GRAB [32]	GraspCVAE [12]	97.10	3.54	2.02	**0.0041**	**2.93**	0.81	2.48
	GraspTTA [12]	**100**	5.05	**1.74**	20.42	2.88	0.93	2.74
	GraspingField [14]	74.80	10.56	3.80	62.46	2.91	3.53	5.83
	Our DVQ-VAE	98.60	**3.18**	2.13	0.15	2.83	**3.64**	**2.45**
Obman [11]	GraspCVAE [12]	99.20	4.32	**1.81**	**0.0040**	2.95	1.50	**2.57**
	GraspTTA [12]	**100**	5.85	2.06	19.70	**2.96**	1.50	3.20
	GraspingField [14]	74.62	10.53	3.81	60.06	2.81	2.33	5.83
	Our DVQ-VAE	99.82	**3.93**	2.70	0.14	2.90	**3.98**	3.07

Table 1: Performance comparison of our proposed DVQ-VAE and the state-of-the-art methods in terms of grasp generation on the widely-adopted benchmarks, *i.e.*, HO-3D [10], FPHA [6], GRAB [32], and Obman [11].

Method	Score
Ours	**3.36**
ContactGen [16]	3.25
GraspTTA [12]	3.23

Table 2: Average scores in human evaluation.

	Contact Ratio (%)↑	Penetration Volume (cm^3)↓	Grasp Disp (cm) ↓
50%Masked	99.60	5.66	2.80
90%Masked	99.70	7.37	2.67

Table 3: Results after masking part of the object point cloud

4.5 Results

Quantitative Results. In experiments, we report the performance in Tab. 1 of our proposed model and the other state-of-the-art methods which are all only trained on Obman [11] and test on HO3D [10], Grab [32], and FPHA [6]. Note that the objects in HO3D, Grab, and FPHA are not present in the Obman training set and are never seen during the training process. As shown in the table, we can find our proposed DVQ-VAE can achieve the lowest penetration and grasp displacement approaches state-of-the-art benchmarks, demonstrating our proposed method allows for fine-grained controlling over touch interactions.

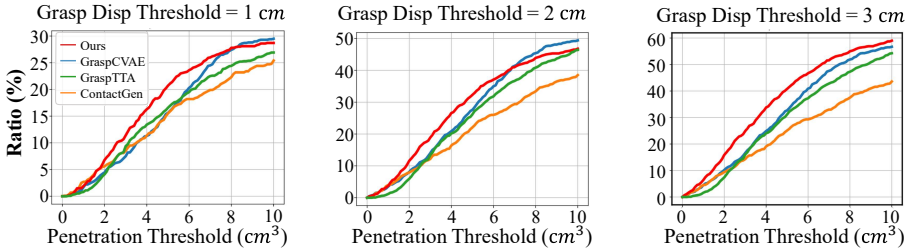

Fig. 3: Performance comparison of our method and other models in high-quality ratio w.r.t the penetration threshold for different models on the HO-3D dataset.

Compared to the model [12] w.r.t quality index, our model shows a remarkable 13.3% relative improvement on HO-3D [10] dataset, 13.7% on FPHA [6] dataset, and 1.2% on GRAB [32] dataset. And competitive results have also been obtained in terms of contact ratio and grasp disp used to assess grasping stability. Although GraspTTA [12] exhibits low grasp disp, its limited grasp diversity results in generating nearly identical grasp for a given object as mentioned in [16]. Moreover, our model achieved a relative improvement of 22.6% in the cluster size metric, suggesting the part-aware decomposed architecture we extended into VQ-VAE [35] allows our model to approach or even surpass the performance of leading methods in terms of grasp diversity. Meanwhile, we measured the ratio of high-quality grasps (both penetration and displacement are below certain thresholds) generated by each model, as shown in Fig. 3 we report the ratio of each model on objects in the HO-3D dataset [10] as the penetration threshold varies from 0 cm^3 to 10 cm^3, and we can clearly see our model outperform the others most of times. These findings indicate the superiority and efficiency of our proposed DVQ-VAE.

Human Evaluation. We also conducted subjective experiments by inviting 10 human participants to evaluate the generated grasps of our DVQ-VAE, ContactGen [16] and GraspTTA [12]. During the testing, given eight objects from the HO-3D [10] dataset, the models randomly generated 4 grasps for each object. Participants rated the grasps with a range from 1 to 5 in terms of *penetration depth*, *grasp stability*, and *naturalness*, and the higher scores indicate the grasp is closer to a real human grasp. We report the results in Fig. 5 and Tab. 2, it is evident that our proposed model produces grasps with the highest 5-point ratio and the lowest 1-point ratios, as well as achieving the highest average human evaluation score.

Robustness. We also tested the performance of our model when only partial point clouds of objects are provided, as shown in Tab. 3. Our model is capable of generating plausible results even with only partial point clouds, because of the powerful prior knowledge embedded in our codebook and autoregressive model. In contrast, ContactGen [16] and GraspTTA [12] can generate grasps only from a fixed number of point clouds.

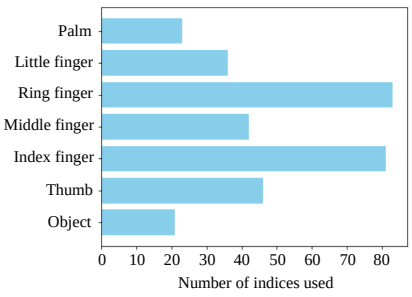

Fig. 4: The number of indices used in each codebook for our DVQ-VAE.

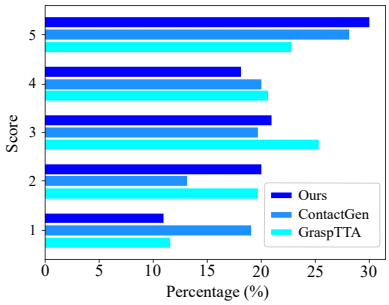

Fig. 5: The percentage of each score of each model in human evaluation.

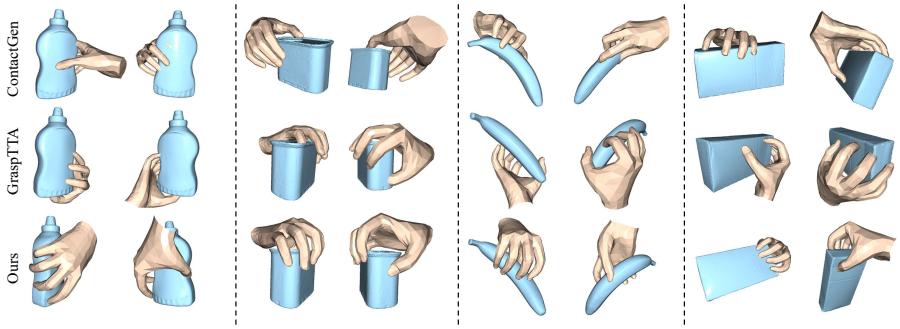

Fig. 6: Qualitative results by comparing our DVQ-VAE with ContactGen [16] and GraspTTA [12] on the HO-3D dataset [10].

Inference Time. As shown in Tab. 1, we achieved a reduction of 99.8%/99.3% time cost compared to ContactGen [16] and GraspTTA [12], respectively, demonstrates our proposed DVQ-VAE obtains faster grasp generation speed compared to methods using optimization methods during the generation period.

Qualitative Results. We visualize the grasps generated for different objects in various datasets, and each grasp is presented from two different perspectives. As Fig. 6 illustrated, we can clearly observe our model generates grasps with fewer penetrations and higher stability than others, showing strong adaptability when applied in such out-of-domain objects. Furthermore, we also visualize the various types of grasps generated for the same object in Fig. 10, and it can be observed that our model can generate grasps with more diversity and different postures. Additionally, we show several visualized failure cases during testing in Fig. 9, we can find our model may generate grasps with insufficient contact when applied to objects with complex geometric shapes. Enhanced representations of objects through the Signed Distance Field (SDF) [21] will be promising to mitigate this challenge.

Method	Penetration Volume (cm^3)↓	Grasp Disp (cm) ↓	Entro-py ↑	Cluster Size↑	Quality Index ↓
VQ-VAE	6.67	7.21	2.82	1.69	7.05
VQ-VAE2	<u>5.18</u>	9.87	**2.96**	<u>4.29</u>	8.46
DVQ-VAE	10.88	4.98	2.94	4.24	6.76
VQ-VAE+Dual-Stage(Two Encoders)	**4.44**	3.61	2.79	1.73	<u>3.86</u>
DVQ-VAE+Dual-Stage(One Encoder)	11.20	4.57	<u>2.95</u>	**4.31**	6.57
DVQ-VAE+Dual-Stage(Reverse)	7.56	<u>2.93</u>	2.79	2.67	4.32
DVQ-VAE+Dual-Stage(Two Encoders)	5.36	**2.75**	2.80	3.84	**3.54**

Table 4: Performance comparison of DVQ-VAE and its variants in the ablation study.

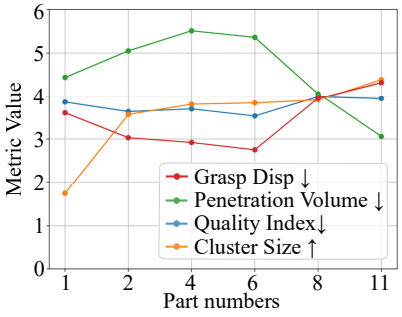

Fig. 7: Performance comparison of different parts used in our proposed DVQ-VAE on the HO-3D [10] dataset. We present the variation trend of evaluation metrics as the number of parts varied.

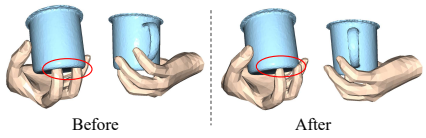

Fig. 8: Comparison of the generated grasp by the model with or without our proposed dual-stage decoding strategy, denoted as "after" and "before", respectively.

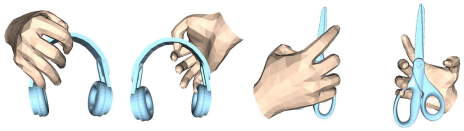

Fig. 9: Visualization examples of our model's failure cases.

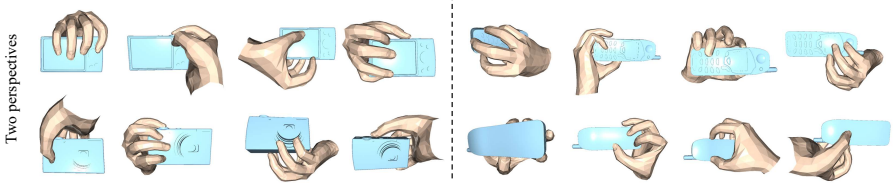

Fig. 10: Diverse grasps generated by our DVQ-VAE for the same object in GRAB [32].

4.6 Ablation Study

Object Encoder We compare the results of encoding the given object with only one object encoder and with our proposed two object encoders in Tab. 4.

We can find our proposed DVQ-VAE with two encoders outperforms the model with only one encoder across most of the metrics, which is attributed to the two encoders that can empower the object representation by decoupling the object's feature into type and pose parts. In addition, Fig. 4 illustrates the distribution of index usage in various codebooks of our DVQ-VAE, it can be seen that objects are clustered into 21 categories based on the latent grasp type. From the usage of indexes, it can be observed that the degrees of freedom of the index finger and the ring finger are higher compared to other parts of the hand, they have more hand posture prototypes.

Part-aware Decomposed Architecture We ablate the decomposed architecture and show the performance in Tab. 4, and we can find the proposed architecture exhibits a relative improvement of 22.4% in the grasp stability compared to vanilla VQ-VAE [35], along with a relative improvement of 151% in cluster size. This implies that our part-aware decomposed architecture not only enhances the diversity of generated grasps but also improves the quality of the generated grasps. Meanwhile, we compared the results in Fig. 7 by utilizing DVQ-VAE with different numbers of parts on HO-3D [10]. It can be observed that dividing the hand into six components yields reasonable results by achieving a good trade-off considering all metrics. Furthermore, in the selection of the backbone network, we also tested VQ-VAE 2.0 [29] structure that incorporates global features. However, as shown in Tab. 4, the results were not satisfactory. Instead, the autoregressive model PixelCNN [22] used in our model can represent latent or implicit global hand features for reconstruction during inference.

Dual-Stage Decoding Strategy. In Tab. 4, we evaluate the effectiveness of our proposed dual-stage decoding strategy. Compared to models without the strategy, our full model can increase the contact ratio, and reduce penetration and displacement, thereby improving the overall grasping quality. Specifically, taking full use of this strategy results in 45.2% and 47.6% relative increases in the quality index of our proposed DVQ-VAE and VQ-VAE [35], respectively. We also tested decoding the position before the posture, but as indicated by the "Dual-Stage (Reverse)" results in the table, the outcome is not promising. Additionally, we also found that the position generation module in our trained DVQ-VAE can optimize the grasps generated by other models by improving the grasping positions. As shown in Fig. 8, the optimized grasps exhibit reduced penetration, validating the effectiveness of our dual-stage decoding strategy.

5 Conclusion

In this paper, we present Decomposed VQ-VAE for human grasp generation, including a part-aware decomposed architecture to account for different components of the hand with latent codebooks, and the dual-stage decoding strategy make the hand posture fit into the position in order. We have shown that each one of these components is important and helps outperform state-of-the-art methods.

Acknowledgements

This work is partly supported by the Funds for the NSFC Project under Grant 62202063, Beijing Natural Science Foundation (L243027), the Innovation Research Group Project of the NSFC under Grant 61921003.

References

1. Boukhayma, A., Bem, R.d., Torr, P.H.: 3d hand shape and pose from images in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10843–10852 (2019)
2. Brahmbhatt, S., Tang, C., Twigg, C.D., Kemp, C.C., Hays, J.: Contactpose: A dataset of grasps with object contact and hand pose. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16. pp. 361–378. Springer (2020)
3. Corona, E., Pumarola, A., Alenya, G., Moreno-Noguer, F., Rogez, G.: Ganhand: Predicting human grasp affordances in multi-object scenes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5031–5041 (2020)
4. Dibra, E., Melchior, S., Balkis, A., Wolf, T., Oztireli, C., Gross, M.: Monocular rgb hand pose inference from unsupervised refinable nets. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1075–1085 (2018)
5. Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3d object reconstruction from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 605–613 (2017)
6. Garcia-Hernando, G., Yuan, S., Baek, S., Kim, T.K.: First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 409–419 (2018)
7. Ge, L., Cai, Y., Weng, J., Yuan, J.: Hand pointnet: 3d hand pose estimation using point sets. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8417–8426 (2018)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
9. Grady, P., Tang, C., Twigg, C.D., Vo, M., Brahmbhatt, S., Kemp, C.C.: Contactopt: Optimizing contact to improve grasps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1471–1481 (2021)
10. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: Honnotate: A method for 3d annotation of hand and object poses. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3196–3206 (2020)
11. Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11807–11816 (2019)
12. Jiang, H., Liu, S., Wang, J., Wang, X.: Hand-object contact consistency reasoning for human grasps generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11107–11116 (2021)

13. Karunratanakul, K., Spurr, A., Fan, Z., Hilliges, O., Tang, S.: A skeleton-driven neural occupancy representation for articulated hands. In: 2021 International Conference on 3D Vision (3DV). pp. 11–21. IEEE (2021)
14. Karunratanakul, K., Yang, J., Zhang, Y., Black, M.J., Muandet, K., Tang, S.: Grasping field: Learning implicit representations for human grasps. In: 2020 International Conference on 3D Vision (3DV). pp. 333–344. IEEE (2020)
15. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
16. Liu, S., Zhou, Y., Yang, J., Gupta, S., Wang, S.: Contactgen: Generative contact modeling for grasp generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20609–20620 (2023)
17. Lv, C., Qi, M., Li, X., Yang, Z., Ma, H.: Sgformer: Semantic graph transformer for point cloud-based 3d scene graph generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4035–4043 (2024)
18. Lv, C., Zhang, S., Tian, Y., Qi, M., Ma, H.: Disentangled counterfactual learning for physical audiovisual commonsense reasoning. Advances in Neural Information Processing Systems **36** (2024)
19. Miller, A.T., Allen, P.K.: Graspit! a versatile simulator for robotic grasping. IEEE Robotics & Automation Magazine **11**(4), 110–122 (2004)
20. Mittal, P., Cheng, Y.C., Singh, M., Tulsiani, S.: Autosdf: Shape priors for 3d completion, reconstruction and generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 306–315 (2022)
21. Oleynikova, H., Millane, A., Taylor, Z., Galceran, E., Nieto, J., Siegwart, R.: Signed distance fields: A natural representation for both mapping and planning. In: RSS 2016 workshop: geometry and beyond-representations, physics, and scene understanding for robotics. University of Michigan (2016)
22. Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al.: Conditional image generation with pixelcnn decoders. Advances in neural information processing systems **29** (2016)
23. Pi, H., Peng, S., Yang, M., Zhou, X., Bao, H.: Hierarchical generation of human-object interactions with diffusion probabilistic models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15061–15073 (2023)
24. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017)
25. Qi, M., Li, W., Yang, Z., Wang, Y., Luo, J.: Attentive relational networks for mapping images to scene graphs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3957–3966 (2019)
26. Qi, M., Qin, J., Li, A., Wang, Y., Luo, J., Van Gool, L.: stagnet: An attentive semantic rnn for group activity recognition. In: Proceedings of the European conference on computer vision (ECCV). pp. 101–117 (2018)
27. Qi, M., Qin, J., Yang, Y., Wang, Y., Luo, J.: Semantics-aware spatial-temporal binaries for cross-modal video retrieval. IEEE Transactions on Image Processing **30**, 2989–3004 (2021)
28. Qi, M., Wang, Y., Li, A., Luo, J.: Stc-gan: Spatio-temporally coupled generative adversarial networks for predictive scene parsing. IEEE Transactions on Image Processing **29**, 5420–5430 (2020)
29. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. Advances in neural information processing systems **32** (2019)
30. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. arXiv preprint arXiv:2201.02610 (2022)

31. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. Advances in neural information processing systems **28** (2015)
32. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: Grab: A dataset of whole-body human grasping of objects. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. pp. 581–600. Springer (2020)
33. Tzionas, D., Ballan, L., Srikantha, A., Aponte, P., Pollefeys, M., Gall, J.: Capturing hands in action using discriminative salient points and physics simulation. International Journal of Computer Vision **118**, 172–193 (2016)
34. Tzionas, D., Gall, J.: 3d object reconstruction from hand-object interactions. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 729–737 (2015)
35. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
36. Wang, C., Wang, L.: Adaptive weight learning for multiple outcome optimization with continuous treatment. arXiv preprint arXiv:2402.11092 (2024)
37. Wang, H., Wang, M., Che, Z., Xu, Z., Qiao, X., Qi, M., Feng, F., Tang, J.: Rgb-depth fusion gan for indoor depth completion. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 6209–6218 (2022)
38. Wang, X., Wu, Y., Zhu, L., Yang, Y.: Symbiotic attention with privileged information for egocentric action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12249–12256 (2020)
39. Wang, X., Zhu, L., Wang, H., Yang, Y.: Interactive prototype learning for egocentric action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8168–8177 (2021)
40. Zheng, Y., Shi, Y., Cui, Y., Zhao, Z., Luo, Z., Zhou, W.: Coop: Decoupling and coupling of whole-body grasping pose generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2163–2173 (2023)
41. Zimmermann, C., Brox, T.: Learning to estimate 3d hand pose from single rgb images. In: Proceedings of the IEEE international conference on computer vision. pp. 4903–4911 (2017)