MaskDexGrasp: Generative Masked Modeling for Part-Aware Dexterous Grasp Synthesis

Binghui Zuo Lin Zhou Haoxuan Xu Jianan Yan Zhipeng Yu Zekai Liu Yangang Wang *

Southeast University, China

Figure 1. Illustration of the proposed MaskDexGrasp. The left side shows the generated dexterous grasps, and the right side shows the diversity and real-world experiments. CVPR is assembled from the palm and fingers to capture the key ideas of our discrete representation.

Abstract

Dexterous grasp generation is a predominant task that enables robots to perform human-level manipulation. However, a dexterous hand always maintains high-dimensional DoF and actuation space, making existing approaches that rely on holistic latent representations difficult to produce high-quality and semantically aligned grasps. In this paper, we propose MaskDexGrasp to address these challenges. We first present a part-aware grasp tokenizer that decomposes dexterous grasps into discrete tokens, facilitating compositional modeling of anatomical dependencies. Building upon this representation, a bidirectional masked grasp transformer is then developed to predict grasp tokens conditioned on object geometry and task description, ensuring coherent grasp generation while allowing fine-grained part-level editing. To facilitate evaluation, we construct a dexterous grasp dataset that comprises 65K grasping instances and 260K richly annotated descriptions covering 11 tasks. Comprehensive experiments demonstrate that our method achieves the state-of-the-art performance. Project page is available at https://binghui-z.github.io/MaskDexGrasp/.

*Corresponding author. E-mail: yangangwang@seu.edu.cn

1. Introduction

Research on dexterous grasp generation [13, 32, 35, 49, 50, 65, 74] chronically stands as a cornerstone in the robotics community, underpinning fine-grained manipulation [7, 8, 29, 64], embodied handover [9, 59, 60], and functional grasp [3, 22, 61, 79, 85]. Unlike parallel-jaw grippers, dexterous hands emulate the anatomical structure of human hands and provide a high degree of freedom, allowing for

intricate interactive precision and adaptive flexibility. Owing to their anthropomorphic capability in handling objects across complex and dynamic environments, advancing dexterous grasp generation is essential for expanding the scope of robotic applications.

However, generating physically feasible and semantically aligned grasps remains a formidable challenge due to its large actuation space. Conventional generative methods predominantly compress grasps into compact latent spaces with VAE-based [25, 26, 34, 36] or diffusion-based [37, 65, 83, 87] models. Although effective for coarse interaction synthesis, such monolithic representations inherently neglect the structured and compositional nature of human-like hands, impeding the learning of fine-grained coordination and diminishing generalization across tasks and object geometries. Furthermore, prior condition-driven grasp generation frameworks solely regard the object geometry as guidance [67, 68, 76], resulting in generated grasps that lack semantic consistency. Although recent advancements [6, 62, 80] attempt to generate stable grasps based on task descriptions, the restricted textual diversity often leads to ambiguous task conditioning, thereby constraining the expressiveness and naturalness of robotic grasping behaviors.

To address these limitations, our insight is to **unify structural decomposition, text-driven conditioning, and controllable grasp generation within a single generative framework**. We observe that dexterous grasps can be hierarchically factorized into part-level hand primitives, each of which performs a distinct yet interdependent functional role. Motivated by this principle, we first introduce a *part-aware grasp tokenizer* that decomposes a given grasp representation into the structured token indices sequence through a vector-quantized variational autoencoder (VQ-VAE) [54]. This formulation transforms the complex manifold of grasps into a discrete space, enabling compositional reasoning and facilitating autoregressive modeling.

Building upon this representation, we further develop a *bidirectional masked grasp transformer* (BMGT) that predicts token indices under diverse conditioning signals, including object geometry and linguistic task semantics. Unlike the standard autoregressive model [12], BMGT employs a bidirectional masked modeling strategy, ensuring the model jointly captures both localized coordination and global interdependence across hand parts. During inference, we adopt an iterative masked sampling scheme equipped with classifier-free guidance [20] to balance fidelity and diversity. Moreover, the discrete representation endows our framework with inherent grasp editability, permitting specific anatomical regions to be resampled without retraining, thereby achieving fine-grained controllability. As illustrated in Fig. 1, our method generates plausible grasps for various objects and realizes satisfactory performance in real-world settings.

In addition, we have also collected a large-scale dexterous grasp dataset called *TDG* to support the construction of our framework. It consolidates heterogeneous grasps from hand-object datasets [24, 72] via a joint-level retargeting solution [44], followed by an energy-based refinement to enforce precise interactions. Each grasp instance is paired with rich descriptions produced from a VLM-assisted captioning protocol [1], providing coherent links between geometric and semantic representations. TDG comprises 64,891 high-quality interaction instances across over 2,296 objects, accompanied by over 259,564 textual annotations spanning 11 task categories.

The main contributions are summarized as follows:

- We present a part-aware grasp tokenizer that discretizes grasp configurations into token indices sequences, enabling compositional reasoning over anatomical parts.
- We develop a bidirectional masked grasp transformer that predicts token indices conditioned on textual semantics and object geometry, while also allowing for flexible editability without retraining.
- We collect a dexterous grasp dataset that offers geometrically precise and semantically grounded interactions, serving as an efficient benchmark for dexterous grasp generation and evaluation.

2. Related Works

Structured Pose Representation. Representing physically feasible and functionally expressive pose configurations is crucial for enabling robots to interact with diverse objects and accomplish human-like complex tasks. Previous approaches typically treated the hand as a monolithic entity, encoding its articulation into continuous pose vectors [2, 5, 48, 84] or low-dimensional latent embeddings [51, 56, 69, 70]. While straightforward, such a holistic representation fails to capture nuanced coordination among fingers, thereby constraining interpretability, modular control, and cross-task generalization. Recognizing this challenge, recent part-based efforts [11, 39] emphasized structured and compositional representations that decompose an articulated entity into meaningful components, reflecting its hierarchical anatomy and functional modularity. Despite its success in universal object part assembly [40, 41, 73, 78], applying such part-level decomposition into articulated hand modeling [10, 66, 81] remains relatively underexplored. In addition, the relevant method SemGrasp [28] discretized grasp to better align with the semantic space, but it still treated the hand as a whole. Motivated by this observation, we advocate decomposing the grasping pose into various semantically meaningful parts to facilitate localized latent learning and modular generation. By modeling the relationships among these parts, it has potential in enhancing the controllability of dexterous grasp generation.

Condition-Driven Grasp Generation. Dexterous grasp-

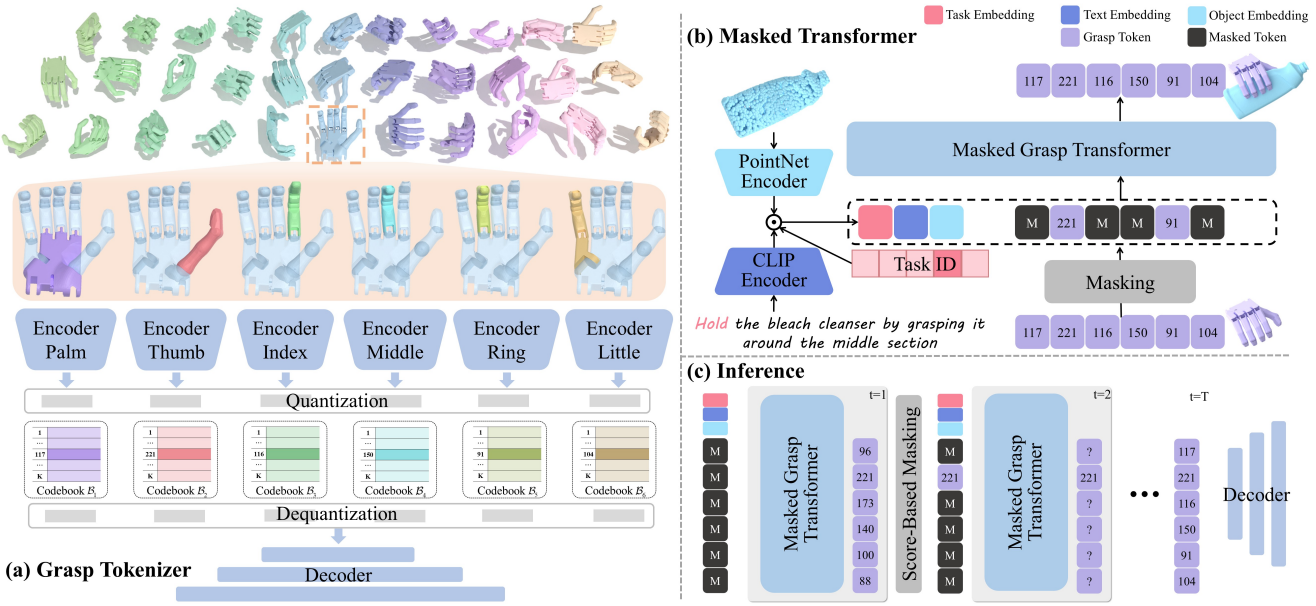

Figure 2. Overview of the proposed MaskDexGrasp. (a) **Grasp Tokenizer** encodes the raw poses into discrete grasp tokens via part-aware VQ-VAE models, which consist of six codebooks (one palm and five fingers). (b) **Masked Transformer** adopts a bidirectional autoregressive manner to learn the probabilistic distributions of these tokens, conditioned on the object geometry and textual description. (c) During **Inference**, the masked transformer progressively and iteratively predicts multiple tokens with high scores from an empty canvas.

ing is inherently context-dependent, as the same object can afford distinct hand configurations depending on task intention, functional demands, or interaction goals. Beyond representing hand structures, effective grasp synthesis increasingly hinges on external conditions to guide generation. Early studies focused on object-driven constraints [26, 34, 82], ensuring grasps were geometrically compatible with the manipulated objects [23, 57, 67] or consistent with the contact priors [25, 34, 37, 87]. However, these methods primarily pursued physical feasibility, neglecting the fine-grained alignment between task intent [24, 72] and object functionality, thus failing to exhibit functional relevance and task awareness. Recently, task-conditioned and intention-aware methods [31, 62, 63, 77] integrated semantic cues and high-level goals, enabling grasps that are both stable and functionally grounded. Despite these advances, most existing methods operated within holistic latent spaces, limiting their ability to disentangle how individual hand components respond to conditioning signals.

Generative Framework Evolution. Prior works typically relied on deterministic mappings between visual observations and hand configurations. The most representative task is CNN-based hand-object reconstruction from RGB images [4, 17–19, 58]. These methods generally adopted either regression-based [18, 33] or optimization-based [14, 53, 71] paradigms, yielding limited diversity in generated grasps. In contrast, probabilistic generative models such as variational autoencoders [27] and diffusion models [21, 47] had been introduced to capture the multimodal distribution of feasible grasps. Nevertheless, most existing approaches [26, 37] encoded the entire grasp holistically into a latent space, failing to exploit the modular nature of hand kinematics. To alleviate this limitation, recent advances in human motion generation [15, 16, 43, 75, 86] demonstrated the effectiveness of discrete latent representations [54] in conjunction with autoregressive mechanisms. By quantizing inputs into discrete codebooks and modeling their dependencies autoregressively, these approaches achieved high-quality and controllable generations. Motivated by these insights, we propose MaskDexGrasp, an autoregressive generative framework that applies part-aware discrete representation for dexterous grasp synthesis.

3. Method

Our objective is to develop a framework for generating dexterous grasps that simultaneously improves grasp quality and accelerates generation efficiency. Towards this goal, as depicted in Fig. 2, our framework consists of two components. First, we present a part-aware grasp tokenizer that decomposes the grasping pose into six anatomical parts, including one palm and five fingers (Sec. 3.1). A bidirectional masked grasp transformer is then designed to autoregressively predict discrete tokens conditioned on object point clouds $\mathcal{O}$ and textual descriptions $\mathcal{T}$ (Sec. 3.2). The inference process, which supports both grasp generation and

finger-level editing, is detailed in Sec. 3.3.

3.1. Part-Aware Grasp Tokenizer

Motivation. Prior methods [26, 36, 52] typically mapped the dexterous grasp into a continuous latent space via VAE models [27]. Despite its compactness, such holistic embedding obscures the structured semantics and inter-finger coordination that are critical for part-level manipulation. In general, a human hand exhibits naturally modular characteristics, where each finger performs distinct yet coordinated functions during grasping.

Formulation. In this work, we utilize the Shadow Hand to parameterize each dexterous grasp, which is represented as $\mathbf{g} = (\theta, \mathbf{R}, \mathbf{t})$ and defined in the object's canonical frame. The $\theta \in \mathbb{R}^{22}$ controls joint angles of the hand model which consists of 22 degrees of freedom, the $\mathbf{R} \in \mathbf{SO}(\mathbf{3})$ denotes the global orientation, and the $\mathbf{t} \in \mathbb{R}^3$ denotes the global translation. Using a pre-defined configuration file (*.xml*), a grasp $\mathbf{g}$ can be transformed to hand surface points $\mathbf{H}$ via the forward kinematics.

Part-Aware Quantization. To overcome the above limitations, we attempt to establish a finger-centric representation and introduce a part-aware grasp quantization strategy, as illustrated in Fig. 2 (a). It decomposes the whole pose into anatomically meaningful subcomponents, while maintaining its structural consistency across different parts. Specifically, we feed grasp $\mathbf{g}$ into the Shadow Hand model and uniformly sample $\mathbf{H} \in \mathbb{R}^{2000\times 3}$ points over the hand surface. These points are subsequently partitioned into N $(N = 6)$ anatomical parts $\{\mathbf{H}_i\}_{i=1}^{N}$, consisting of one palm and five fingers (*i.e.* , thumb, index, middle, ring, and little finger). Each part is encoded into a latent embedding $\mathbf{z}_i$ through a corresponding encoder $\mathcal{E}_i$:

$$\mathbf{z}_i = \mathcal{E}_i(\mathbf{H}_i), \quad i \in \{1, \dots, 6\}. \tag{1}$$

For the extracted latent embedding $\mathbf{z}_i$, it is quantized by a learnable codebook $\mathcal{B}_i = \{\mathbf{b}_i^k\}_{k=1}^{K}$ containing K entries, where the goal is to find the closest Euclidean distance between the embedding $\mathbf{z}_i$ and codebook element $\mathbf{b}_i^k$ through the following equation.

$$\hat{\mathbf{z}}_i = \mathbf{b}_i^{\mathbf{s}_i}, \ where\ \mathbf{s}_i = \arg\min_k \|\mathbf{z}_i - \mathbf{b}_i^k\|_2. \tag{2}$$

By this way, we can discretize latent embeddings $\mathbf{z} = \{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_N\}$ into corresponding codebook entries $\hat{\mathbf{z}} = \{\hat{\mathbf{z}}_1, \hat{\mathbf{z}}_2, \cdots, \hat{\mathbf{z}}_N\}$ and obtain a sequence of indices $\mathbf{S} = \{\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_N\}$ which serves as the compositional representation of the input grasp $\mathbf{g}$. The target grasp configuration $\hat{\mathbf{g}}$ is reconstructed via a shared decoder $\hat{\mathbf{g}} = \mathcal{D}(\hat{\mathbf{z}})$.

Training Objective. We train the grasp tokenizer following the standard VQ-VAE objective, which comprises reconstruction and commitment losses as follows:

$$\mathcal{L}_{\text{VQ}} = \mathcal{L}_{rec.} + \sum_{i=1}^{6} \Big(\|\text{sg}[\mathbf{z}_i] - \hat{\mathbf{z}}_i\|_2^2 + \beta \|\mathbf{z}_i - \text{sg}[\hat{\mathbf{z}}_i]\|_2^2 \Big), \tag{3}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operation and β is a hyper-parameter to balance the codebook commitment term. For the reconstruction loss $\mathcal{L}_{rec.}$, it enforces the VQ-VAE to reconstruct plausible hand grasp $\hat{\mathbf{g}}$, containing constraints of the pose parameters and surface points.

$$\mathcal{L}_{rec.} = \|\hat{\mathbf{g}} - \mathbf{g}\|_2 + \sum_{i=1}^{N} \Big(\hat{\mathbf{H}}_i - \mathbf{H}_i \Big). \tag{4}$$

3.2. Bidirectional Masked Grasp Transformer

Motivation. With a learned part-aware tokenizer, a sequence of indices $\mathbf{S} = \{\mathbf{s}_i\}_{i=1}^{6}$ can be projected back to codebook entries to reconstruct the target grasp $\hat{\mathbf{g}}$. Although this discrete formulation naturally facilitates generative sequence modeling, a key challenge lies in how to learn a conditional distribution that could produce part-aware tokens and maintain flexibility for high-fidelity generation. To this end, we develop a *bidirectional masked grasp transformer* (BMGT), a conditional generative model that operates entirely in the discrete latent space. As shown in Fig. 2 (b), it learns to autoregressively predict grasp tokens, while leveraging object geometry and task semantics as complementary conditions.

Formulation. Given a token sequence $\mathbf{S}$, the objective is to model a conditional distribution $p(\mathbf{S} \mid \mathbf{C})$, where $\mathbf{C}$ encapsulates generative conditions. Different from conventional autoregressive models that predict tokens in a next-index order, BMGT employs a bidirectional masked modeling scheme [16, 30], allowing tokens to be inferred from both preceding and succeeding contextual information. Concretely, a random subset of tokens in $\mathbf{S}$ is masked to form the corrupted sequence $\mathbf{S_M}$, and the network learns to recover these masked elements guided by the condition $\mathbf{C}$.

$$\mathcal{L} = - \mathop{\mathbb{E}}_{\mathbf{S}\in p(\mathbf{S})} \left[\sum_{\forall i \in [1,N]} \log p(\mathbf{s}_i \mid \mathbf{S_M},\ \mathbf{C}) \right]. \tag{5}$$

Task-Conditioned Masked Transformer. To enable semantically guided synthesis, we serve the task embedding `[TASK]`, the text embedding $\mathbf{t}$, as well as the object point cloud embedding $\mathbf{o}$ as the inputs of BMGT. Specifically, a learnable embedding layer learns the task embedding `[TASK]` from a task ID that refers to an action type, a pre-trained CLIP model $\mathcal{F}_{\mathcal{T}}$ [45] transforms the textual instruction $\mathcal{T}$ into a dense text embedding $\mathbf{t} = \mathcal{F}_{\mathcal{T}}(\mathcal{T})$, and a PointNet-based encoder $\mathcal{F}_{\mathcal{O}}$ extracts the object geometric embedding $\mathbf{o} = \mathcal{F}_{\mathcal{O}}(\mathcal{O})$ from the object point clouds $\mathcal{O}$. These embeddings are mapped into a unified latent space and concatenated together to form the conditional vector $\mathbf{C} = [\texttt{[TASK]}; \mathbf{t}; \mathbf{o}]$. Furthermore, the discretized token indices $\mathbf{S}$ are randomly masked out and replaced by learnable `[MASK]` tokens to obtain $\mathbf{S_M}$. Practically, we adopt the same masking schedule $\gamma(\tau) = \cos\left(\frac{\pi\tau}{2}\right)$ as the related

work [16]. Both $\mathbf{S_M}$ along with the conditional vector $\mathbf{C}$ are fed into a BMGT model to predict masked tokens through a bidirectional autoregressive mechanism. The `[END]` token is not considered, as the sequence length is fixed. We optimize the transformer by minimizing the negative log-likelihood of the predicted tokens.

$$\mathcal{L}_{\text{AR}} = -\sum_{i=1}^{N} \log p(\mathbf{s}_i \mid \mathbf{S_M}, (\texttt{[TASK]}; \mathbf{t}; \mathbf{o})). \quad (6)$$

3.3. Inference and Controllable Generation

Iterative Masked Sampling. We utilize an iterative masked sampling procedure that relies on a score-based masking scheme to generate grasps in inference. As shown in Fig. 2 (c), starting from a fully masked token sequence $\mathbf{S}^{(0)} = \{\texttt{[MASK]}_1, \texttt{[MASK]}_2, \ldots, \texttt{[MASK]}_N\}$, the model progressively predicts and replaces masked tokens with the most probable candidates over T iterations. At each iteration t, BMGT predicts grasp tokens at masked locations along with their probabilities. For those tokens with the lowest $\lceil \gamma\left(\frac{t}{T}\right) \cdot N \rceil$ confidence, we mask out again and concatenate them with others to estimate expected tokens at the next $t+1$ iteration, until t reaches T. Finally, all the predicted tokens are mapped back to codebook embeddings and decoded to $\hat{\mathbf{g}}$ through the VQ-VAE decoder.

Classifier-Free Guidance. We also apply *classifier-free guidance* [20] to enhance the fidelity and diversity in the generated results. Concretely, BMGT is trained with conditional dropout, denoting that the conditioning signal $\mathbf{C}$ is randomly omitted with a probability $p_{\text{uncond}} = 10\%$. While during inference, the final logits ω_g are formed by shifting the conditional logits ω_c away from the unconditional logits ω_u with a guidance scale s:

$$\omega_g = (1+s) \cdot \omega_c - s \cdot \omega_u. \quad (7)$$

Controllable Grasp Editing. Benefiting from the part-aware quantization and bidirectional causal mask, our model provides a natural extension for localized grasp editing. In detail, given an estimated token sequence $\hat{\mathbf{S}}$, the relevant tokens corresponding to particular fingers that require editing can be re-masked and re-sampled under the modified condition $\mathbf{C}'$. Formally, for an editable hand region $\Omega \subseteq \{1, \ldots, N\}$, we fix the unmasked context $\mathbf{S}_{\overline{\Omega}}$ and only update the masked tokens:

$$\tilde{\mathbf{S}}_{\Omega} \sim p(\mathbf{S}_{\Omega} \mid \mathbf{S}_{\overline{\Omega}}, \mathbf{C}'). \quad (8)$$

The updated token sequence $\tilde{\mathbf{S}}$ is subsequently decoded by the VQ-VAE decoder to yield the edited grasp $\tilde{\mathbf{g}}$. This formulation offers a direct and interpretable method for modifying high-level grasp semantics through discrete tokens, thereby enabling part-specific modifications. Crucially, this operation requires no retraining, delivering a potential application for interactive grasp controllability and task-conditioned adaptation.

4. Dataset

Existing datasets struggle to balance both the data scale and semantic richness, particularly in the absence of aligned task categories. To support the construction and evaluation of our framework, we collect a large-scale dexterous grasp dataset, named *TDG*, which includes 64,891 grasping annotations, 259,564 textual descriptions, 2,296 object models, and 11 task categories.

Grasping Pose Retargeting. We collect paired hand-object interactions from existing datasets [24, 72]. A primary challenge in building our dataset arises from the heterogeneous pose representations in both datasets. AffordPose [24] represents grasps through the DoF of a MANO-based kinematic structure, while OakShape [72] represents grasps using MANO parameters [48]. The discrepancy between the two representations makes direct integration infeasible. To resolve this issue, we adopt a joint-level retargeting strategy, which performs the retargeting process in two steps. First, given the joint positions $\mathbf{J}_k$ from both datasets, we compute an initial retargeted grasp $\mathbf{g}_0$ following [44], which optimizes dexterous grasp parameters based on the joint alignment. However, such purely geometric alignment neglects the constraints between the hand and the object. To ensure physically plausible and semantically consistent interaction, we further introduce an optimization stage that refines $\mathbf{g}_0$ by minimizing the energy function. More implementation details can be found in the appendix.

Language Annotation via VLM. With the assistance of VLM (*qwen-vl-max* [1] is leveraged in our experiments), we further design a language annotation protocol to enrich our dataset with semantic groundings. Specifically, for each hand-object interaction, we render images from a virtual camera viewpoint that visually depict the interaction status. We then employ the VLM to produce diverse language annotations using both the rendered image and the concise description as inputs, where the latter provides corresponding task and object category information. To further enhance linguistic richness, we prompt the model to produce four temporal variants of the same grasp description, covering basic, present, progressive, and passive tenses.

Dataset Split. Consistent with the original structure in [24, 72], TDG is divided into two subsets, where the *Subset 1* refers to AffordPose [24] and *Subset 2* refers to OakShape [72]. In practice, we perform all experiments on these two sets respectively. Each subset is divided into train/test splits with the percentage of 90%/10%.

5. Experiments

5.1. Implementation Details

Our framework is implemented using PyTorch [42] and built in two stages. For the grasp tokenizer, the codebook size is set to 256×512. Each encoder $\mathcal{E}_i$ employs a Point-

Table 1. Quantitative comparisons on *Subset 1* and *Subset 2* demonstrate that our method achieves superior performance across almost all metrics. The ■ denotes the best results, ■ denotes the secondary, and ■ denotes the tertiary.

Methods	*Subset 1*					*Subset 2*				
	Suc. ↑	*Q1* ↑	*Pen.* ↓	H_{mean} ↑	H_{std} ↓	*Suc.*↑	*Q1* ↑	*Pen.* ↓	H_{mean} ↑	H_{std} ↓
DGTR [67]	27.57	0.039	0.343	3.000	0.708	29.18	0.058	0.279	1.823	0.367
DexGYS [62]	30.61	0.038	0.471	3.611	0.572	38.54	0.070	0.494	3.078	0.411
SceneDiffuser [23]	35.46	0.045	0.471	3.878	0.460	53.31	0.092	0.489	3.218	0.365
UGG [37]	33.63	0.039	0.350	3.835	0.496	43.31	0.069	0.442	3.448	0.346
DexGraspAnything [83]	46.54	0.042	0.376	4.207	0.377	58.90	0.125	0.477	3.662	0.253
Ours	44.68	0.048	0.340	3.876	0.421	75.16	0.126	0.413	3.922	0.406

Net backbone followed by linear layers to map the part surface points into latent spaces, and the decoder consists of fully connected layers with ReLU activations. Following [16, 75], to avoid codebook collapse [46], the exponential moving average (EMA) and codebook reset (Code Reset) is applied. For the BMGT, we employ 9 transformer layers [55], with 16 heads and an embedding dimension of 512. We warm up the learning rate to $2e^{-4}$ after 2000 iterations for both stages. The grasp tokenizer is trained for 200 epochs with a batch size of 256, and the BMGT is trained for 500 epochs with a batch size of 128. We perform both training processes on a single NVIDIA GeForce RTX 4070Ti Super GPU, costing 7 and 18 hours respectively.

5.2. Experimental Setups

Metrics. Five metrics are reported to present comprehensive evaluations from the aspects of the grasp quality and diversity. Those are, 1) Success rate (%) (*Suc.*) measures the proportion of successful grasps. A grasping pose is considered successful if it maintains stable interaction in at least one of the six gravity directions and also satisfies a maximal penetration depth of less than 5mm. We use the IsaacGym simulator [38] with the same settings used in [37, 57]. 2) *Q1* reflects grasp stability, where we set the contact threshold to 1cm and set the penetration threshold to 5mm following [57]. 3) Maximal penetration depth (cm) (*Pen.*) computes the maximal penetration depth from the object point cloud to the hand mesh. 4) H_{mean} and H_{std} quantify the diversity of joint angle distributions. A higher mean entropy indicates more diverse grasp distributions, while a lower standard deviation suggests consistent diversity across all samples. 5) Inference time (s) refers to the mean time required to generate a single pose.

Baselines. Prior works that follow a generative manner are selected as our baselines, including DGTR [67], DexGYS [62], SceneDiffuser [23], UGG [37], and DexGraspAnything [83]. All methods are retrained and evaluated on our dataset for fair comparisons.

Figure 3. Visualization of the generated diverse grasps, where two rows respectively represent the model trained on two subsets.

5.3. Comparisons with SOTA Methods

The quantitative results on both subsets are summarized in Tab. 1. As observed, our method consistently outperforms other baselines across most metrics. The higher success rate and lower penetration depth reflect stable and high-quality generations, which suggest the effectiveness of our framework for dexterous grasp synthesis. Notably, the better penetration performance of DGTR may be caused by unstable grasping, as its *Suc.* is poorer than others. Although diffusion-based approach [83] exhibits comparable performance in diversity, a characteristic where VQVAE underperforms, our method is still convincing after balancing the improvement in generation quality.

Fig. 4 visualizes the qualitative comparisons, showcasing that our method produces anatomically coherent and semantically meaningful grasps. For instance, MaskDexGrasp predicts stable interactions for target objects and confirms that the generated grasps are aligned with the given task, such as twist and use. In contrast, other methods [23, 37, 67] generate grasps without requiring textual descriptions, ensuring only reasonable grasping, but neglecting specific task awareness. In Fig. 3, we illustrate more diverse generations from our method, showing that our method achieves visually acceptable diversity.

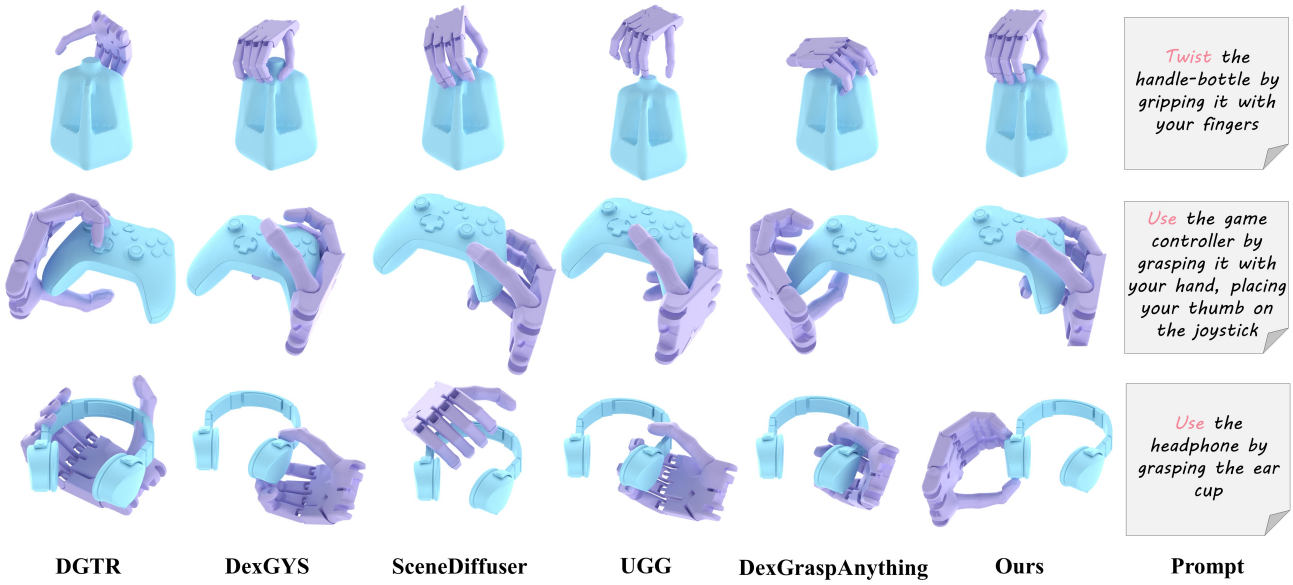

Figure 4. Qualitative comparisons of the generated dexterous grasps from different methods.

Figure 5. Visualization of the edited grasps, where purple indicates the original generation and orange denotes the grasp after editing.

Table 2. Computational cost of model parameters and inference times for a single generation.

Method	[67]	[62]	[23]	[37]	[83]	Ours
Param ↓	3.85	23.14	22.98	67.03	159.68	71.29
Time ↓	0.284	0.202	1.130	3.236	4.417	0.033

5.4. Inference Speed and Editability

We report computational costs through model parameters and inference times for generating a single grasp in Tab. 2. All experiments are performed on a single NVIDIA GeForce RTX 4070Ti Super GPU with a batch size of 1. Compared to baselines, our autoregressive framework receives superior performance in inference time. This is because the diffusion-based method typically requires at least 50 denoising steps, which is significantly more than our sampling frequency. Moreover, we conduct experiments to validate the editability of MaskDexGrasp, which is defined as finger-level modifications. By selecting specific hand part indices, the model can resample grasp tokens and alter finger poses without disturbing the overall configuration or retraining. Fig. 5 shows editing results across one or two fingers, highlighting the flexibility of our framework.

5.5. Ablation Studies

Our ablations involve the part-aware tokenizer and the grasp transformer. 1) For *the architecture of tokenizer*, we have conducted ablation studies to explore the importance of part-aware structure and placed it with a vanilla VQ-VAE [54]. It means that all hand surface points are compressed into the latent space via a single encoder-decoder network. From the degraded results in Row-1 of Tab. 3, we observe that the adopted architecture exhibits a significant improvement across all metrics. 2) For *the dimension of codebook*, we recognize that the trainable parameters in codebook impact generation performance, and it is crucial to adopt an appropriate codebook dimension for balancing generation accuracy and computational cost. Therefore, the dimension of $128 \times 256, 512 \times 1024$ are included in Tab. 3. Detailed relationships among codebook dimension, model parameter, and penetration depth are shown in Fig. 6 (left), demonstrating our settings are reasonable and effectively mitigate the collapse issue faced by VQ-VAE. 3) For *the number of iteration* during inference, it directly influences the generation speed and quality. As reported in Tab. 3, although increasing the number of iteration leads to higher performance, the gains become negligible once a certain number is reached. At the same time, the occupied time required for generation will also increase. More ablation studies on the number

Table 3. Ablation studies of key components in our framework, including the settings of codebook and masked grasp transformer.

Methods	*Subset 1*					*Subset 2*				
	Suc. ↑	*Q1* ↑	*Pen.* ↓	H_{mean} ↑	H_{std} ↓	*Suc.*↑	*Q1* ↑	*Pen.* ↓	H_{mean} ↑	H_{std} ↓
architecture of tokenizer										
w/ vanilla	32.55	0.039	0.433	3.426	0.538	50.63	0.093	0.498	3.320	0.427
dimension of codebook										
128×256	39.87	0.042	0.402	3.503	0.489	61.64	0.114	0.482	3.426	0.453
512×1024	43.92	0.046	0.367	3.751	0.434	72.57	0.122	0.410	3.818	0.411
number of iteration										
iteration 3	45.65	0.048	0.355	3.867	0.418	74.47	0.128	0.464	3.911	0.408
iteration 5	44.95	0.049	0.355	3.869	0.420	74.66	0.128	0.409	3.909	0.407
Ours	44.68	0.048	0.340	3.876	0.421	75.16	0.126	0.413	3.922	0.406

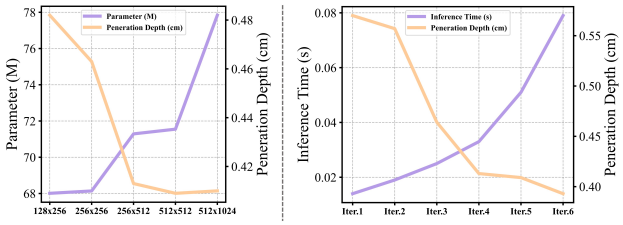

Figure 6. Ablations of the codebook dimension and iteration number, which are performed on *Subset 2*.

of iteration, inference time, and penetration depth are illustrated in Fig. 6 (right), which suggests our scheme balances the inference time and generation quality.

5.6. Real-World Experiments

To confirm the adaptability of our framework in practical applications, we conduct real-world experiments and deploy the model on a physical robotics system. As illustrated in Fig. 7, it consists of an XArm7 robot arm and a Freedom five-fingered dexterous hand. For each target grasp, we first move the arm to a pre-grasp position to avoid unnecessary collisions between the robot and the object, and then move the arm to the hand root position to execute the predicted pose. A grasp is regarded as successful when the object can be stably lifted by the hand without slipping or dropping. These real-world experiments denote that our method achieves robust and satisfactory grasping across different objects. Additional details and dynamic demonstrations can be found in our appendix.

6. Conclusion

In this paper, we introduce a novel generative framework for part-aware dexterous grasp synthesis. By representing high-dimensional dexterous grasp as discrete tokens through a part-aware grasp tokenizer, our method enables compositional reasoning across anatomical parts. We further design a bidirectional masked grasp transformer that is jointly conditioned on the object geometry and textual description to predict the token indices sequence autoregressively, allowing coherent grasp generation with fine-grained part-level editing. Additionally, we construct a TDG dataset to support comprehensive training and evaluation, which offers large-scale interactions with rich linguistic annotations. Extensive experiments demonstrate that our approach achieves the state-of-the-art performance in dexterous grasp generation and provides explicit controllability.

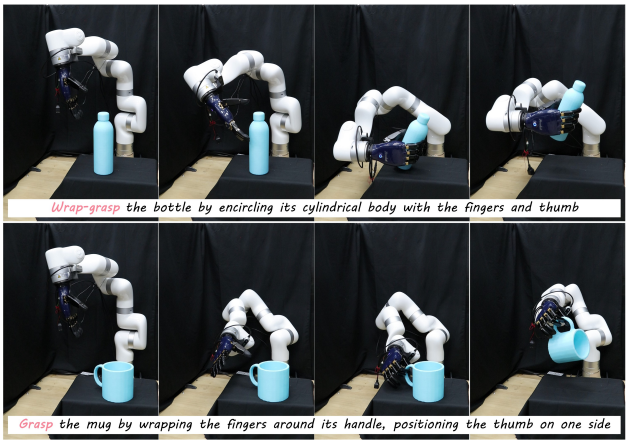

Figure 7. Real-world experiments of our method, where we select four key frames for each successful grasp.

Limitations. Despite the promising results, our framework currently relies on finite discrete spaces and offline tokenization, which may limit its diversity and adaptability to continuous control. Future work will extend this paradigm to dynamic grasp generation, as well as multi-step manipulation, thereby bridging the gap between practical human skills and robotics intelligent operations.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China (No. 62076061), in part by the Natural Science Foundation of Jiangsu Province (No. BK20220127), in part by the Postgraduate Research&Practice Innovation Program of Jiangsu Province (No. SJCX25_0080).

References

[1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 2, 5

[2] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10843–10852, 2019. 2

[3] Samarth Brahmbhatt, Ankur Handa, James Hays, and Dieter Fox. Contactgrasp: Functional multi-finger grasp synthesis from contact. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2386–2393. IEEE, 2019. 1

[4] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *European Conference on Computer Vision*, pages 361–378. Springer, 2020. 3

[5] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 666–682, 2018. 2

[6] Xiaoyun Chang and Yi Sun. Text2grasp: Grasp synthesis by text prompts of object grasping parts. *arXiv preprint arXiv:2404.15189*, 2024. 2

[7] Jiayi Chen, Yubin Ke, Lin Peng, and He Wang. Dexonomy: Synthesizing all dexterous grasp types in a grasp taxonomy. *Robotics: Science and Systems*, 2025. 1

[8] Yuanpei Chen, Yiran Geng, Fangwei Zhong, Jiaming Ji, Jiechuang Jiang, Zongqing Lu, Hao Dong, and Yaodong Yang. Bi-dexhands: Towards human-level bimanual dexterous manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):2804–2818, 2023. 1

[9] Sammy Christen, Wei Yang, Claudia Pérez-D'Arpino, Otmar Hilliges, Dieter Fox, and Yu-Wei Chao. Learning human-to-robot handovers from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9654–9664, 2023. 1

[10] Enric Corona, Tomas Hodan, Minh Vo, Francesc Moreno-Noguer, Chris Sweeney, Richard Newcombe, and Lingni Ma. Lisa: Learning implicit shape and appearance of hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20533–20543, 2022. 2

[11] Boyang Deng, John P Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Nasa neural articulated shape approximation. In *European Conference on Computer Vision*, pages 612–628. Springer, 2020. 2

[12] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2

[13] Carlo Ferrari, John Canny, et al. Planning optimal grasps. In *Proceedings., 1992 IEEE International Conference on Robotics and Automation, 1992.*, pages 2290–2295. IEEE, 1992. 1

[14] Patrick Grady, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C Kemp. Contactopt: Optimizing contact to improve grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1471–1481, 2021. 3

[15] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5152–5161, 2022. 3

[16] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024. 3, 4, 5, 6

[17] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3196–3206, 2020. 3

[18] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11807–11816, 2019. 3

[19] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 571–580, 2020. 3

[20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 5

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[22] Linyi Huang, Hui Zhang, Zijian Wu, Sammy Christen, and Jie Song. Fungrasp: functional grasping for diverse dexterous hands. *IEEE Robotics and Automation Letters*, 2025. 1

[23] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16750–16761, 2023. 3, 6, 7

[24] Juntao Jian, Xiuping Liu, Manyi Li, Ruizhen Hu, and Jian Liu. Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In *Proceedings*

of the IEEE/CVF International Conference on Computer Vision, pages 14713–14724, 2023. 2, 3, 5, 1

[25] Juntao Jian, Xiuping Liu, Zixuan Chen, Manyi Li, Jian Liu, and Ruizhen Hu. G-dexgrasp: Generalizable dexterous grasping synthesis via part-aware prior retrieval and prior-assisted generation. *arXiv preprint arXiv:2503.19457*, 2025. 2, 3

[26] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11107–11116, 2021. 2, 3, 4

[27] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3, 4

[28] Kailin Li, Jingbo Wang, Lixin Yang, Cewu Lu, and Bo Dai. Semgrasp: Semantic grasp generation via language aligned discretization. In *European Conference on Computer Vision*, pages 109–127. Springer, 2024. 2

[29] Kailin Li, Puhao Li, Tengyu Liu, Yuyang Li, and Siyuan Huang. Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6991–7003, 2025. 1

[30] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37:56424–56445, 2024. 4

[31] An-Lun Liu, Yu-Wei Chao, and Yi-Ting Chen. Task-oriented human grasp synthesis via context-and task-aware diffusers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10375–10385, 2025. 3

[32] Min Liu, Zherong Pan, Kai Xu, Kanishka Ganguly, and Dinesh Manocha. Generating grasp poses for a high-dof gripper using neural networks. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1518–1525. IEEE, 2019. 1

[33] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14687–14697, 2021. 3

[34] Shaowei Liu, Yang Zhou, Jimei Yang, Saurabh Gupta, and Shenlong Wang. Contactgen: Generative contact modeling for grasp generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20609–20620, 2023. 2, 3

[35] Tengyu Liu, Zeyu Liu, Ziyuan Jiao, Yixin Zhu, and Song-Chun Zhu. Synthesizing diverse and physically stable grasps with arbitrary hand structures using differentiable force closure estimator. *IEEE Robotics and Automation Letters*, 7(1): 470–477, 2021. 1

[36] Yumeng Liu, Yaxun Yang, Youzhuo Wang, Xiaofei Wu, Jiamin Wang, Yichen Yao, Sören Schwertfeger, Sibei Yang, Wenping Wang, Jingyi Yu, et al. Realdex: Towards human-like grasping for robotic dexterous hand. *arXiv preprint arXiv:2402.13853*, 2024. 2, 4

[37] Jiaxin Lu, Hao Kang, Haoxiang Li, Bo Liu, Yiding Yang, Qixing Huang, and Gang Hua. Ugg: Unified generative grasping. In *European Conference on Computer Vision*, pages 414–433. Springer, 2024. 2, 3, 6, 7

[38] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021. 6

[39] Marko Mihajlovic, Shunsuke Saito, Aayush Bansal, Michael Zollhoefer, and Siyu Tang. Coap: Compositional articulated occupancy of people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13201–13210, 2022. 2

[40] Paritosh Mittal, Yen-Chi Cheng, Maneesh Singh, and Shubham Tulsiani. Autosdf: Shape priors for 3d completion, reconstruction and generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 306–315, 2022. 2

[41] Abhinav Narayan, Rajendra Nagar, and Shanmuganathan Raman. Rgl-net: A recurrent graph learning framework for progressive part assembly. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 78–87, 2022. 2

[42] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[43] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1546–1555, 2024. 3

[44] Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, and Dieter Fox. Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system. In *Robotics: Science and Systems*, 2023. 2, 5

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 4

[46] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 6

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[48] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 2, 5, 1

[49] Yanming Shao and Chenxi Xiao. Bimanual grasp synthesis for dexterous robot hands. *IEEE Robotics and Automation Letters*, 2024. 1

[50] Qijin She, Shishun Zhang, Yunfan Ye, Ruizhen Hu, and Kai Xu. Learning cross-hand policies of high-dof reaching and grasping. In *European Conference on Computer Vision*, pages 269–285. Springer, 2024. 1

[51] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 89–98, 2018. 2

[52] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pages 581–600. Springer, 2020. 4

[53] Tze Ho Elden Tse, Zhongqun Zhang, Kwang In Kim, Ales Leonardis, Feng Zheng, and Hyung Jin Chang. S 2 contact: Graph-based network for 3d hand-object contact estimation with semi-supervised learning. In *European Conference on Computer Vision*, pages 568–584. Springer, 2022. 3

[54] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2, 3, 7

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 6

[56] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Crossing nets: Combining gans and vaes with a shared latent space for hand pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 680–689, 2017. 2

[57] Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. *arXiv preprint arXiv:2210.02697*, 2022. 3, 6

[58] Yinqiao Wang, Hao Xu, Pheng-Ann Heng, and Chi-Wing Fu. Unihope: A unified approach for hand-only and hand-object pose estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12231–12241, 2025. 3

[59] Youzhuo Wang, Jiayi Ye, Chuyang Xiao, Yiming Zhong, Heng Tao, Hang Yu, Yumeng Liu, Jingyi Yu, and Yuexin Ma. Dexh2r: A benchmark for dynamic dexterous grasping in human-to-robot handover. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 1

[60] Zifan Wang, Junyu Chen, Ziqing Chen, Pengwei Xie, Rui Chen, and Li Yi. Genh2r: learning generalizable human-to-robot handover via scalable simulation demonstration and imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16362–16372, 2024. 1

[61] Wei Wei, Peng Wang, Sizhe Wang, Yongkang Luo, Wanyi Li, Daheng Li, Yayu Huang, and Haonan Duan. Learning human-like functional grasping for multifinger hands from few demonstrations. *IEEE Transactions on Robotics*, 40: 3897–3916, 2024. 1

[62] Yi-Lin Wei, Jian-Jian Jiang, Chengyi Xing, Xian-Tuo Tan, Xiao-Ming Wu, Hao Li, Mark Cutkosky, and Wei-Shi Zheng. Grasp as you say: Language-guided dexterous grasp generation. *Advances in Neural Information Processing Systems*, 37:46881–46907, 2024. 2, 3, 6, 7

[63] Yi-Lin Wei, Mu Lin, Yuhao Lin, Jian-Jian Jiang, Xiao-Ming Wu, Ling-An Zeng, and Wei-Shi Zheng. Afforddexgrasp: Open-set language-guided dexterous grasp with generalizable-instructive affordance. *arXiv preprint arXiv:2503.07360*, 2025. 3

[64] Zhenyu Wei, Zhixuan Xu, Jingxiang Guo, Yiwen Hou, Chongkai Gao, Zhehao Cai, Jiayu Luo, and Lin Shao. D (r, o) grasp: A unified representation of robot and object interaction for cross-embodiment dexterous grasping. *arXiv preprint arXiv:2410.01702*, 2024. 1

[65] Zehang Weng, Haofei Lu, Danica Kragic, and Jens Lundell. Dexdiffuser: Generating dexterous grasps with diffusion models. *IEEE Robotics and Automation Letters*, 2024. 1, 2

[66] Wei Xie, Zimeng Zhao, Shiying Li, Binghui Zuo, and Yangang Wang. Nonrigid object contact estimation with regional unwrapping transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9342–9351, 2023. 2

[67] Guo-Hao Xu, Yi-Lin Wei, Dian Zheng, Xiao-Ming Wu, and Wei-Shi Zheng. Dexterous grasp transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17933–17942, 2024. 2, 3, 6, 7

[68] Yinzhen Xu, Weikang Wan, Jialiang Zhang, Haoran Liu, Zikang Shan, Hao Shen, Ruicheng Wang, Haoran Geng, Yijia Weng, Jiayi Chen, et al. Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4737–4746, 2023. 2

[69] Linlin Yang and Angela Yao. Disentangling latent hands for image synthesis and pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9877–9886, 2019. 2

[70] Linlin Yang, Shile Li, Dongheui Lee, and Angela Yao. Aligning latent spaces for 3d hand pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2335–2343, 2019. 2

[71] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Cpf: Learning a contact potential field to model the hand-object interaction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11097–11106, 2021. 3

[72] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20953–20962, 2022. 2, 3, 5, 1

[73] Guanqi Zhan, Qingnan Fan, Kaichun Mo, Lin Shao, Baoquan Chen, Leonidas J Guibas, Hao Dong, et al. Generative 3d part assembly via dynamic graph learning. *Advances in Neural Information Processing Systems*, 33:6315–6326, 2020. 2

[74] Hui Zhang, Sammy Christen, Zicong Fan, Otmar Hilliges, and Jie Song. Graspxl: Generating grasping motions for di-

verse objects at scale. In *European Conference on Computer Vision*, pages 386–403. Springer, 2024. 1

[75] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14730–14740, 2023. 3, 6

[76] Jialiang Zhang, Haoran Liu, Danshi Li, XinQiang Yu, Haoran Geng, Yufei Ding, Jiayi Chen, and He Wang. Dexgraspnet 2.0: Learning generative dexterous grasping in large-scale synthetic cluttered scenes. In *8th Annual Conference on Robot Learning*, 2024. 2

[77] Jieyi Zhang, Wenqiang Xu, Zhenjun Yu, Pengfei Xie, Tutian Tang, and Cewu Lu. Dextog: Learning task-oriented dexterous grasp with language condition. *IEEE Robotics and Automation Letters*, 2024. 3

[78] Rufeng Zhang, Tao Kong, Weihao Wang, Xuan Han, and Mingyu You. 3d part assembly generation with instance encoded transformer. *IEEE Robotics and Automation Letters*, 7(4):9051–9058, 2022. 2

[79] Yibiao Zhang, Jinglue Hang, Tianqiang Zhu, Xiangbo Lin, Rina Wu, Wanli Peng, Dongying Tian, and Yi Sun. Functionalgrasp: Learning functional grasp for robots via semantic hand-object representation. *IEEE Robotics and Automation Letters*, 8(5):3094–3101, 2023. 1

[80] Zhongqun Zhang, Hengfei Wang, Ziwei Yu, Yihua Cheng, Angela Yao, and Hyung Jin Chang. Nl2contact: Natural language guided 3d hand-object contact modeling with diffusion model. In *European Conference on Computer Vision*, pages 284–300. Springer, 2024. 2

[81] Zhe Zhao, Mengshi Qi, and Huadong Ma. Decomposed vector-quantized variational autoencoder for human grasp generation. In *European Conference on Computer Vision*, pages 447–463. Springer, 2024. 2

[82] Juntian Zheng, Qingyuan Zheng, Lixing Fang, Yun Liu, and Li Yi. Cams: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 585–594, 2023. 3

[83] Yiming Zhong, Qi Jiang, Jingyi Yu, and Yuexin Ma. Dexgrasp anything: Towards universal robotic dexterous grasping with physics awareness. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22584–22594, 2025. 2, 6, 7, 3

[84] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5346–5355, 2020. 2

[85] Tianqiang Zhu, Rina Wu, Jinglue Hang, Xiangbo Lin, and Yi Sun. Toward human-like grasp: Functional grasp by dexterous robotic hand via object-hand semantic representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12521–12534, 2023. 1

[86] Qiran Zou, Shangyuan Yuan, Shian Du, Yu Wang, Chang Liu, Yi Xu, Jie Chen, and Xiangyang Ji. Parco: Part-coordinating text-to-motion synthesis. In *European Conference on Computer Vision*, pages 126–143. Springer, 2024. 3

[87] Binghui Zuo, Zimeng Zhao, Wenqian Sun, Xiaohan Yuan, Zhipeng Yu, and Yangang Wang. Graspdiff: Grasping generation for hand-object interaction with multimodal guided diffusion. *IEEE Transactions on Visualization and Computer Graphics*, 31(9):5642–5654, 2025. 2, 3