

TaxoGrasp: Taxonomy-Guided Human Grasp Synthesis with Sparse Contact Constraint

Haitian Liu¹, Yin Wang³, Zhiying Leng¹, Kanglei Zhou², Yan Wang⁵,
Frederick W. B. Li⁴, and Xiaohui Liang^{1,5*}

¹ State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, China

² Department of Psychological and Cognitive Sciences, Tsinghua University, China

³ Artificial Intelligence Research Institute, Central University of Finance and Economics, China

⁴ Department of Computer Science, Durham University, United Kingdom

⁵ Zhongguancun Laboratory, China
liang_xiaohui@buaa.edu.cn

Abstract. We tackle controllable human grasp synthesis under sparse user-specified *thumb-contact constraints*. While existing methods use contact as conditioning, they lack explicit semantic grounding, yielding weak contact-intent coupling, poor constraint sensitivity, and pose drift. We propose **TaxoGrasp**, a framework threading Feix grasp taxonomy as a *persistent intent bottleneck* throughout: (1) intent inference via supervised taxonomy prediction from object geometry and thumb-contact constraint, (2) diffusion synthesis conditioned on taxonomy templates, and (3) anchor-optimized grounding preserving taxonomy structure. We augment HOGraspNet with patch-level thumb-contact annotations and introduce a standardized evaluation protocol with three metric categories. Experiments demonstrate improved intent consistency, physical plausibility, and constraint adherence.

Keywords: Hand Grasp Synthesis · Hand-Object Interaction · Human Grasp Taxonomy

1 Introduction

Controllable human grasp synthesis aims to generate physically feasible hand poses that satisfy user-specified conditions [9, 15, 21, 23, 35, 39, 43]. It plays a fundamental role in robot manipulation, AR/VR interaction, and 3D content creation, where users expect to directly control how objects are grasped [7, 12, 31, 46, 47]. In realistic interactive scenarios, language can express intent but is often difficult to ground into precise geometric constraints [38]; conversely, richly specified geometric contact conditions are informative yet impractical. As a result, a sparse yet critical geometric contact constraint has emerged as a more desirable control signal [37]. However, sparse constraints often introduce ambiguity in intent specification [21, 28, 43]. How to accurately infer grasp intent under such sparse geometric cues is the key to enabling reliable and controllable grasp synthesis, as illustrated in Fig. 1(a).

* Corresponding author.

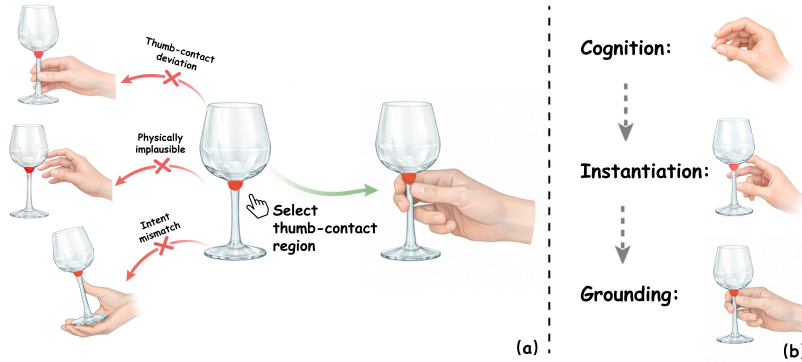


Fig. 1: Illustration of the TaxoGrasp setting. Given an object and a user-specified thumb-contact region on its surface (highlighted), our goal is to synthesize a physically feasible and semantically intent-consistent human grasp pose. **(a)** One correct grasp and three representative failure cases: thumb-contact deviation, physical implausibility, and intent mismatch. **(b)** Cognition-Instantiation-Grounding three-stage illustration.

Recent advances in grasp synthesis have increasingly adopted a contact-centric control paradigm, wherein object-aware contact cues are either predicted or explicitly specified first. Subsequently, full hand configurations are generated conditioned on these contact cues [21, 28, 29, 44]. This contact-first, pose-second formulation has proven effective in improving the local physical plausibility of grasp synthesis. Contact cues provide direct geometric grounding, which aids in enforcing contact consistency during both the generation and post-hoc refinement stages [14, 20]. However, the contact-centric paradigm can become fragile when the conditioning signal is sparse, as local contact constraints alone are insufficient to determine the global hand configuration. A typical example of this challenge can be seen in RegionGrasp [37], which uses the thumb-contact region as a condition for grasp synthesis. We observe that while this approach generates multiple grasp poses that satisfy the sparse contact constraints, it may result in solutions that are semantically misaligned with user intentions. Moreover, the absence of explicit intention modeling may yield physically suboptimal grasps, as illustrated in Fig. 3. This highlights a significant limitation of the contact-centric approach: the difficulty in inferring grasping intentions and generating intention-consistent grasps from sparse geometric constraints alone.

Inferring user intent from sparse contact cues faces three key challenges: (1) many-to-many ambiguity between local contacts and global grasp mode, making the problem ill-posed [24, 28]; (2) lack of quantifiable supervision beyond geometric losses, as intent combines object geometry, contact cues, and human grasping priors [30]; and (3) optimization conflict where contact-driven refinement improves physical plausibility but often destroys semantic organization, preventing grasps that satisfy both user-expected organization AND geometric executability [19, 27].

We address these challenges with TaxoGrasp, a taxonomy-first framework for interactive grasp synthesis (Fig. 1(b)). Our key insight introduces the Feix taxonomy [11] as a *discrete semantic bottleneck* by **first converting ill-posed intent inference into supervised classification** conditioned on object geometry and thumb-contact region, then **threading this taxonomy as a persistent organizational prior through all three stages**: Cognition (intent inference), Instantiation (diffusion conditioning), and Grounding (optimization guidance). This unified treatment ensures that human intent is faithfully and physically realized through a cohesive intent-to-execution pipeline.

We also construct an interactive evaluation benchmark on top of HOGraspNet [8], a dataset that provides real grasp taxonomy labels, by augmenting each object with patch-level thumb-contact region annotations, enabling reproducible evaluation under sparse thumb-contact control. We introduce a comprehensive evaluation protocol that jointly quantifies interactive feasibility, generation quality, and condition compliance. Extensive qualitative and quantitative experiments demonstrate that, compared with existing contact-centric baselines, TaxoGrasp generates more intent-consistent grasps with improved physical plausibility. Our main contributions include:

- Reformulate sparse-contact grasp synthesis as supervised taxonomy inference conditioned on object geometry and thumb region, explicitly resolving contact-semantics decoupling under inherent ambiguity.
- Propose TaxoGrasp, a taxonomy-first pipeline comprising taxonomy inference, conditioned diffusion synthesis, and novel anchor-level taxonomy templates that partition contact/non-contact regions for optimization guidance.
- Augment HOGraspNet with patch-level thumb-contact annotations and introduce a standardized evaluation protocol assessing condition consistency, interaction quality, and generation fidelity.

2 Related Work

Controllable hand grasp synthesis aims to generate physically plausible, diverse hand configurations while respecting user-specified task conditions, typically by combining conditional generative modeling with feasibility priors and contact-/collision-aware regularization or refinement. Existing formulations can be broadly grouped into (i) natural-language control and (ii) geometric/contact-based control. Language-conditioned methods treat free-form text as a flexible specification of grasp intent (e.g., grasp type, target part, action semantics) and learn conditional generators that align synthesized grasps with linguistic semantics, often introducing intermediate structured representations or additional constraints to mitigate the intrinsic ambiguity of language and to maintain contact realism [4, 5, 23, 39, 43]. In contrast, geometric/contact-based approaches employ explicit, verifiable conditions, such as contact maps, semantic contact patterns, as control signals, and steer generation toward satisfying these constraints while enforcing physical plausibility through consistency objectives and optimization-based refinement [14, 20, 21, 24, 28, 37]. Both paradigms emphasize compliance

with explicit control signals, including semantic alignment for language or geometric satisfaction for contact inputs, under feasibility priors. In contrast, our method infers user grasp intent from sparse, underdetermined cues rather than merely enforcing those cues as hard constraints.

Human grasp taxonomies offer a discrete, interpretable semantic space that abstracts diverse hand configurations and contact strategies. The GRASP taxonomy of Feix et al. categorizes stable one-handed static grasps into a finite set of types using structured criteria (e.g., opposition, virtual-finger assignment, and power-precision properties) [11]. HOGraspNet further supports taxonomy-based supervision with large-scale human hand-object interactions annotated by Feix labels and paired 3D hand/object data [8]. Discreteness in this semantic space has also been used to structure generation. Dexonomy and OmniDexVLG perform Feix-type-conditioned synthesis and refinement for controllable robotic grasp generation [6, 41]. Recent human grasp synthesis methods similarly introduce discrete semantics or codes with staged decoding to capture fine-grained interactions [23, 45]. However, prior work typically treats taxonomy as a conditioning interface or as motivation for discrete structure, rather than a persistently inferred representation reusable across tasks. We instead supervise taxonomy inference under the Feix taxonomy, yielding reusable grasp templates that enable taxonomy-conditioned synthesis and taxonomy-guided alignment.

3 Method

3.1 Overview

Given an object O , we allow the user to select a thumb-contact region C_{th} from the N pre-partitioned surface regions (Sec. 4.2). Our goal is to synthesize a MANO [34] grasp pose H^* that (i) places the thumb on C_{th} , (ii) is geometrically executable on O , and (iii) matches the user’s intended grasp mode under thumb-contact control. TaxoGrasp bridges discrete grasp semantics and continuous contact feasibility via a three-stage pipeline (Fig. 2).

Stage I: Cognition (Human Grasp Template Reasoning). From (O, C_{th}) , we infer a high-level grasp template $\hat{y}$ in the Feix taxonomy [11] and produce a unified thumb-aware condition embedding $\mathbf{c}$ that summarizes object geometry together with the thumb constraint.

Stage II: Instantiation (Template-Conditioned Grasp Synthesis). Conditioned on $(\mathbf{c}, \hat{y})$, we sample an initial continuous MANO pose H_0 using a Conditional-Diffusion-Transformer [32, 42], modeling $p(H_0 \mid \mathbf{c}, \hat{y})$. This stage provides an initialization close to the contact-feasible manifold, which is critical for subsequent geometric refinement.

Stage III: Grounding (Taxonomy-Guided Contact Alignment). Given $(O, C_{\text{th}}, \hat{y}, H_0)$, we ground the grasp by instantiating a taxonomy-specific contact structure over hand anchors (contact-relevant surface keypoints introduced in [40]), partitioning them into contact and non-contact sets. We refine H_0 by attracting contact anchors to the object surface (with thumb anchors restricted

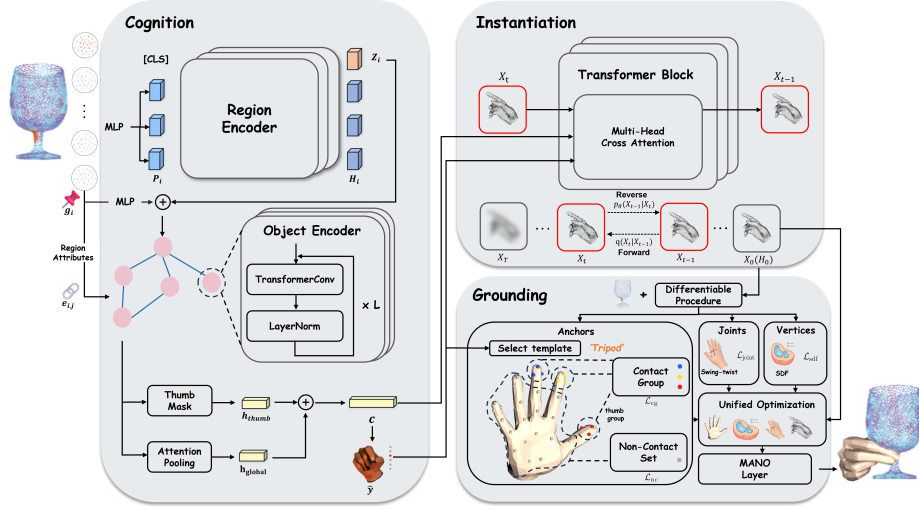


Fig. 2: TaxoGrasp framework. Given an object mesh O and a user-specified thumb-contact region C_{th} , **Cognition**(Sec. 3.2) infers a Feix grasp [11] template $\hat{y}$ and a thumb-aware condition embedding c . **Instantiation**(Sec. 3.3) samples an initial MANO [34] grasp $H_0 \sim p(H | c, \hat{y})$ with a conditional diffusion transformer [32, 42]. **Grounding**(Sec. 3.4) grounds the grasp by taxonomy-guided contact alignment, refining H_0 into an executable pose H^* .

to C_{th}) while repelling non-contact anchors to preserve functional semantics. The optimized pose is returned as H^* .

3.2 Cognition: Human Grasp Template Reasoning

Stage I performs template reasoning and thumb-aware conditional representation from a user-specified thumb-contact region. Given object geometry O and the selected thumb-contact region C_{th} , it predicts a discrete Feix grasp taxonomy [11] $\hat{y}$ that specifies the global grasp organization, and produces a compact condition embedding c that summarizes the object geometry with the thumb constraint.

Region encoder. We encode each surface region with a shared Transformer and use the [CLS] token as a fixed-length region descriptor. For region i with point features $\mathcal{P}_i = \{\mathbf{p}_{i,j}\}_{j=1}^{|\mathcal{P}_i|}$, where $\mathbf{p}_{i,j}$ are per-point features extracted by a shared MLP. We project points to the token space and prepend a learnable [CLS] token:

$$\mathbf{X}_i = [\mathbf{t}_{cls}; \mathbf{p}_{i,1} \mathbf{W}_p; \dots; \mathbf{p}_{i,|\mathcal{P}_i|} \mathbf{W}_p] \in \mathbb{R}^{(1+|\mathcal{P}_i|) \times d}. \quad (1)$$

A shared Transformer encoder $\text{Enc}(\cdot)$ outputs region tokens

$$\mathbf{H}_i = \text{Enc}(\mathbf{X}_i), \quad \mathbf{z}_i = \mathbf{H}_i[cls] \in \mathbb{R}^d. \quad (2)$$

Object encoder. Our surface partitioning strategy (Sec. 4.2) naturally induces a region-level surface adjacency, which we treat as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to propagate local context over neighboring regions for thumb-aware reasoning. Each node i carries a geometric attribute vector $\mathbf{g}_i$ (e.g., patch center/normal statistics and a thumb indicator), and each edge (i, j) carries a relation feature $\mathbf{e}_{ij}$ capturing pairwise surface geometry (e.g., center distance and normal agreement). We initialize node states by fusing region descriptors and attributes:

$$\mathbf{h}_i^{(0)} = \phi(\mathbf{W}_h [\mathbf{z}_i; \mathbf{W}_g \mathbf{g}_i]), \quad (3)$$

and apply L layers of edge-conditioned graph Transformer updates:

$$\mathbf{h}_i^{(l+1)} = \text{LN}\left(\mathbf{h}_i^{(l)} + \phi\left(\text{GraphT}(\mathbf{h}_i^{(l)}, \{\mathbf{h}_j^{(l)}\}_{j \in \mathcal{N}(i)}, \{\mathbf{e}_{ij}\})\right)\right). \quad (4)$$

In practice, $\text{GraphT}(\cdot)$ is implemented as multi-head graph attention with edge attributes (PyG `TransformerConv` [13]), where $\mathbf{e}_{ij}$ conditions attention-based message passing when aggregating neighboring regions, $\text{LN}(\cdot)$ denotes the LayerNorm [1] and $\phi(\cdot)$ denotes the GELU [16] activation function.

Thumb-aware readout and taxonomy prediction. Let $\tau_i \in \{0, 1\}$ denote a one-hot thumb mask over regions, where $\tau_i = 1$ iff region i equals the selected thumb-contact region C_{th} . We read out a global object feature via attention pooling and extract a thumb-focused feature via masked aggregation, forming the unified condition embedding:

$$\mathbf{c} = \left[\sum_{i=1}^N \tau_i \mathbf{h}_i^{(L)} ; \text{AttPool}(\{\mathbf{h}_i^{(L)}\}) \right]. \quad (5)$$

where $\text{AttPool}(\cdot)$ denotes attention pooling using learned attention weights. We then predict taxonomy logits from $\mathbf{c}$ using a MLP and train Stage I with cross-entropy supervision; at inference time, we output $(\hat{y}, \mathbf{c})$ as the discrete template and thumb-aware condition for subsequent stages.

3.3 Instantiation: Template-Conditioned Grasp Synthesis

Stage II instantiates the discrete template into a continuous MANO [34] grasp by sampling from a conditional diffusion model (Fig. 2). Given the thumb-aware condition embedding $\mathbf{c}$ and the inferred Feix taxonomy [11] $\hat{y}$ from Stage I (Sec. 3.2), we model the distribution of MANO parameters conditioned on both continuous geometry cues and the discrete grasp organization as $p(\mathbf{H}_0 \mid \mathbf{c}, \hat{y})$, to generate an initial grasp for downstream grounding. We denote the clean MANO parameters by $\mathbf{x}_0$, which corresponds to the initial grasp H_0 . Unlike instantiating a fixed taxonomy prototype, conditional sampling captures intra-class variation while remaining anchored to the intended grasp family, which is crucial since the same grasp type may correspond to diverse hand configurations under different object geometries and thumb placements.

We formulate grasp synthesis as a diffusion process in MANO parameter space. Let $\mathbf{x}_0 \in \mathbb{R}^D$ denote the target MANO parameters and $\mathbf{x}_t$ the noisy sample at diffusion step t . Following the standard DDPM [18] formulation, $\mathbf{x}_t$ is obtained by progressively adding Gaussian noise to $\mathbf{x}_0$ under a predefined variance schedule. A Transformer-based denoiser f_θ predicts the clean parameters conditioned on $\mathbf{c}$ and a Feix taxonomy feature:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad \hat{\mathbf{x}}_0 = f_\theta(\mathbf{x}_t, \mathbf{c}, \mathbf{z}_{\text{feix}}(\hat{y}), t). \quad (6)$$

To aggregate heterogeneous conditions, we tokenize pose, object, and taxonomy cues and perform conditional fusion via cross-attention at each Transformer layer. At every timestep, we construct a pose token from $(\mathbf{x}_t, t)$, an object token from $\mathbf{c}$, and a Feix taxonomy token from $\mathbf{z}_{\text{feix}}(\hat{y})$. The pose token serves as the query and cross-attends to the condition tokens, producing a conditioned representation used for diffusion denoising.

The Feix taxonomy feature $\mathbf{z}_{\text{feix}}(\cdot)$ is constructed from a single canonical pose prototype (a medoid) for each Feix category, selected from the training set. This design provides a structured organization prior that is more informative than a discrete label alone (Detailed definitions in supplementary materials).

During training, we use the ground-truth taxonomy label y to construct $\mathbf{z}_{\text{feix}}(y)$ and optimize the standard diffusion regression objective:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{x}_0, t} \left[\left\| f_\theta(\mathbf{x}_t, \mathbf{c}, \mathbf{z}_{\text{feix}}(y), t) - \mathbf{x}_0 \right\|_2^2 \right]. \quad (7)$$

At inference time, we replace y with the predicted $\hat{y}$ and iteratively denoise from Gaussian noise to obtain the initial grasp H_0 .

3.4 Grounding: Taxonomy-Guided Contact Alignment

Although Stages I–II (Secs. 3.2 and 3.3) provide a semantically coherent grasp template and an instantiated MANO [34] pose, the synthesized grasp is not guaranteed to be geometrically executable on the object. Stage III therefore *grounds* the grasp by solving a differentiable optimization problem: starting from the synthesized pose H_0 , we refine it into an executable pose H^* under explicit *soft* geometric and anatomical constraints, while preserving the functional semantics implied by the inferred taxonomy $\hat{y}$ (Fig. 2).

We adopt the differentiable hand anchors defined in [40], i.e., a sparse set of surface keypoints $\{\mathbf{a}_k(H)\}_{k=1}^K$ on the MANO mesh. Given the inferred taxonomy label $\hat{y}$, all anchors are partitioned into contact groups $\{\mathcal{G}_y^{(m)}\}_{m=1}^{M_{\hat{y}}}$ and a non-contact set $\mathcal{N}_{\hat{y}}$ according to a taxonomy-specific *contact template*, which we manually annotate. Specifically, the thumb contact group $\mathcal{G}_{\hat{y}}^{(m_{\text{th}})}$ consists of the anchors located on the distal phalanx of the thumb.

We encourage functional contact by requiring that for each contact group, *at least one* anchor establishes contact with its target. For the thumb contact group (denoted as m_{th}), the target is the center $\mathbf{c}_{\text{th}}$ of the thumb-contact region

C_{th} ; for all other groups, the target is the object surface ∂O . Specifically, we define the target distance $d_k^{(m)}(H)$ for anchor k in group m as:

$$d_k^{(m)}(H) = \begin{cases} \|\mathbf{a}_k(H) - \mathbf{c}_{\text{th}}\|_2, & m = m_{\text{th}}, \\ |s_O(\mathbf{a}_k(H))|, & m \neq m_{\text{th}}, \end{cases} \quad (8)$$

where $s_O(\cdot)$ is the object’s signed distance function (SDF) [36]. This objective is formulated via a soft-min aggregation within each group:

$$\mathcal{L}_{\text{cg}}(H, \hat{y}) = \frac{1}{M_{\hat{y}}} \sum_{m=1}^{M_{\hat{y}}} \text{softmin}_{k \in \mathcal{G}_{\hat{y}}^{(m)}} \left(d_k^{(m)}(H) \right), \quad (9)$$

where $\text{softmin}(\{u_k\}) = -\frac{1}{\alpha} \log \sum_k \exp(-\alpha u_k)$ with $\alpha = 20$. For the non-contact set $\mathcal{N}_{\hat{y}}$, we enforce a clearance margin $m_{\text{nc}} = 2$ mm to preserve functional semantics:

$$\mathcal{L}_{\text{nc}}(H, \hat{y}) = \frac{1}{|\mathcal{N}_{\hat{y}}|} \sum_{k \in \mathcal{N}_{\hat{y}}} \left(\max(0, m_{\text{nc}} - s_O(\mathbf{a}_k(H))) \right)^2. \quad (10)$$

In addition, we incorporate vertex- and joint-level constraints to ensure physical plausibility. At the vertex level, we penalize interpenetration by minimizing $\mathcal{L}_{\text{vertex}}$, which targets MANO vertices $\mathbf{v}$ where $s_O(\mathbf{v}) < 0$. At the joint level, we apply $\mathcal{L}_{\text{joint}}$, which employs swing-twist decomposition [2, 40] to penalize anatomically implausible rotations and explicitly penalizes inter-finger intersections. Detailed formulations are provided in the supplementary materials. The final executable pose H^* is obtained by minimizing the joint objective:

$$\begin{aligned} \mathcal{L}(H) = & \lambda_{\text{vertex}} \mathcal{L}_{\text{vertex}}(H) + \lambda_{\text{joint}} \mathcal{L}_{\text{joint}}(H) \\ & + \lambda_{\text{cg}} \mathcal{L}_{\text{cg}}(H, \hat{y}) + \lambda_{\text{nc}} \mathcal{L}_{\text{nc}}(H, \hat{y}) + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}(H, H_0) \end{aligned} \quad (11)$$

where $\mathcal{L}_{\text{reg}}(H, H_0)$ is a regularization term penalizing excessive deviation from the initial synthesis.

4 Experiments

4.1 Implementation Details

For **Cognition**, the batch size is set to 64, the number of training epochs is set to 150, the learning rate is 0.001; the patch-level Transformer uses $d_{\text{model_pc}} = 256$, the geometry branch dimension is set to $d_{\text{geom}} = 128$, the hidden size is 256, and the number of layers is 6. For **Instantiation**, we employ a Transformer-based Diffusion generator with latent dimension 256, time embedding dimension 512, 6 Transformer layers, 4 attention heads, FFN dimension 512, and dropout 0.1; in the diffusion process, the variances β_t are predefined to linearly spread from 0.0001 to 0.02, and the total number of noising steps is set to $T = 1000$, while

classifier-free guidance is adopted at inference with a guidance scale of 1.5. For **Grounding**, we refine the grasp by contact alignment on a precomputed SDF grid, where the SDF refinement runs for 100 iterations with a learning rate of 0.01, and we set $\lambda_{\text{vertex}} = 1$, $\lambda_{\text{joint}} = 0.5$, $\lambda_{\text{cg}} = 0.5$, $\lambda_{\text{nc}} = 0.1$, $\lambda_{\text{reg}} = 2$; the non-contact margin is set to 2 mm, and a softmin aggregation is used with $\alpha = 20$. All training and inference are conducted on a single NVIDIA RTX 4060 Ti GPU.

4.2 Datasets

All experiments are conducted on **HOGraspNet** [8], which contains RGB-D recordings from **four** synchronized viewpoints of **99** participants manipulating **30** everyday objects, with large-scale annotated frames and corresponding 3D hand-object information. HOGraspNet is built upon the Feix grasp taxonomy [11]: it covers **33** everyday grasp types and further merges visually/geometrically redundant categories into **28** classes while keeping indices aligned with the Feix taxonomy. Following the authors’ released code, we use all 15,987 labeled samples from `labeling_data` and construct stratified train/val/test splits by grasp category. We use HOGraspNet since our pipeline requires explicit grasp taxonomy labels, which are unavailable in most real hand-object interaction datasets.

Region Partition and Thumb-contact Annotation. We assume each object surface is partitioned into N regions so that the user-specified thumb contact can be expressed as a discrete region index. For each object mesh, we construct a region graph by partitioning the surface into regions of comparable physical size, making the representation largely independent of mesh resolution. Specifically, we choose the number of regions based on the mesh surface area and a target region diameter, and obtain the partition with a surface *Capacity-Constrained Centroidal Voronoi Tessellation (CCVT)* [3, 22]. In our implementation, we set the target region diameter to 2 cm, and determine the number of regions based on the mesh surface area A as $N \approx A / (\pi(1 \text{ cm})^2)$; hence, N scales with object size. Each region is defined as the set of surface samples assigned to its CCVT cell; since the partition is induced by surface sampling and CCVT enforces approximate balance, the number of points per region is variable. We define region adjacency from mesh connectivity: two regions are adjacent if there exists a mesh edge whose two incident triangles are assigned to different regions. Finally, to derive the thumb-contact region from the original grasp annotation, we transform the MANO [34] hand mesh into the object coordinate frame, compute a predefined thumb anchor [40], and label the region whose center is closest to this anchor as the thumb-contact region. (Detailed definitions in the supplementary materials).

4.3 Metrics

We evaluate with three metric groups, i.e., **condition consistency**, **interaction quality**, and **generation quality**, each aligned with a core requirement of

interactive grasp synthesis under a sparse thumb-region cue: stable constraint adherence, mode-consistent synthesis, and physically feasible grounding. Detailed metric definitions are provided in the supplementary materials.

Condition consistency measures whether the model consistently enforces the specified thumb-contact constraint across samples by reporting **Hit Rate**, i.e., the fraction of generated grasps whose thumb contact is consistent with the user-designated region C_{th} ; higher values indicate more reliable satisfaction of the conditioning cue.

Interaction quality evaluates physical executability, reflecting whether the synthesized grasps are feasible when interacting with the target object. We report **Contact Area (CA)** to quantify contact support, **Maximum Penetration Depth (PD(max))**, **Mean Penetration Depth (PD(mean))**, and **Penetration Volume (PV)** to characterize interpenetration severity, and **Simulation Displacement (SIM)** to measure stability in a physics simulator [10] (smaller is more stable), jointly assessing whether generated grasps can stably grasp the object.

Generation quality characterizes grasp-organization consistency by measuring the distribution gap between generated samples and real grasps: if the model captures the grasp organization implied by the thumb cue, the generated set should align with real data in category-level statistics and overall distribution. Following [8, 33], we use the same frozen grasp taxonomy classifier as a common evaluator and report **TDD** (taxonomy distribution divergence), where $q \in \Delta^C$ is the empirical histogram of ground-truth taxonomy labels and $p \in \Delta^C$ is the histogram of evaluator-predicted labels for generated grasps, and report $TDD = 10 \times JSD(p|q)$ for numerical readability, with standard ϵ -smoothing and renormalization [26]; We also report **cFID**, the Fréchet distance [17] between real and generated grasp feature distributions extracted from the penultimate layer of the evaluator. Lower values indicate better distribution-level consistency.

4.4 Comparison

We quantitatively compare **TaxoGrasp** with representative approaches that are most relevant to our setting, including **GraspTTA** [20], **ContactGen** [28], **ClickDiff** [24], and **RegionGrasp** [37]. Tab. 1 and Fig. 4 summarize the quantitative results, and Fig. 3 provides qualitative comparisons.

From the perspective of *interaction quality*, our method yields the largest **contact area** (CA), indicating more effective hand-object support while respecting the user-specified thumb constraint. Crucially, the improved contact is not obtained by pushing the hand into the object: TaxoGrasp also attains the lowest **penetration depth** (PD_{max}/PD_{mean}) and a substantially smaller **penetration volume** (PV), reflecting cleaner hand-object geometry with reduced interpenetration. This geometric feasibility further translates into physics stability, where TaxoGrasp obtains the smallest **simulation displacement** (SIM), suggesting more stable and executable grasps under gravity and contact dynamics. Fig. 3 corroborates these trends: across different objects and thumb-contact conditions, our results exhibit more appropriate contact establishment with visibly fewer penetrations. In terms of *generation quality*, TaxoGrasp achieves the

Table 1: Quantitative comparison of state-of-the-art methods on the HOGraspNet [8] dataset. The best and second-best results are shown in bold and underlined.

Method	Interaction					Generation	
	CA $\uparrow$	PD(max) $\downarrow$	PD(mean) $\downarrow$	PV $\downarrow$	SIM $\downarrow$	cFID $\downarrow$	TDD $\downarrow$
GraspTTA [20]	<u>113.4</u>	0.561	0.236	8.085	12.89	11.75	0.953
ContactGen [28]	106.2	0.494	0.218	5.437	12.59	6.754	0.502
ClickDiff [24]	102.5	<u>0.392</u>	<u>0.188</u>	<u>3.420</u>	11.13	9.149	0.655
RegionGrasp [37]	110.8	0.486	0.212	4.309	<u>11.11</u>	<u>2.968</u>	<u>0.435</u>
Ours	120.1	0.340	0.131	1.265	10.72	1.714	0.223

lowest **cFID**, indicating that the distribution of generated grasps is closest to real data in the evaluator feature space. Moreover, it also attains the lowest **taxonomy distribution divergence** (TDD). This result suggests that our approach better preserves organization-level semantic stability, producing grasps whose taxonomy statistics align more closely with those of real human grasps. As illustrated in Fig. 3 using the wine-glass example, the two thumb-contact conditions implicitly correspond to two distinct grasp organizations; our method correctly recovers the intended “holding” and “clinking” grasps, whereas both baselines fail to exhibit a clear organization-dependent distinction. Finally, regarding *condition consistency*, TaxoGrasp achieves the highest **Hit Rate** (Fig. 4), demonstrating the most reliable adherence to the user-designated thumb-contact region. In Fig. 3, the orange marker indicates the specified thumb-contact region, and our method shows the strongest compliance with this constraint.

Taken together, by introducing the Feix taxonomy [11] as a semantic interface and maintaining taxonomy consistency throughout our pipeline, TaxoGrasp successfully uncovers the grasp intent implicitly conveyed by a sparse thumb-contact cue and synthesizes grasps that better match user intention, remain structurally stable, and are geometrically executable.

4.5 Component Analysis

To comprehensively validate the design of TaxoGrasp, we conduct ablation studies following the three-stage structure of our pipeline. Quantitative results are reported in Tab. 2 and Fig. 4, and representative qualitative comparisons are shown in Fig. 6.

Stage I: Cognition. We empirically validate the reliability of the taxonomy classifier, which achieves **72.8%** Top-1 accuracy and **99.8%** Top-3 accuracy on the test split. While Top-1 is not perfect, the high Top-3 indicates that the correct (or a near-equivalent) grasp organization is almost always within a small candidate set. This makes many Top-1 deviations *tolerable* for downstream synthesis, mainly due to two characteristics of region-level intent: (i)

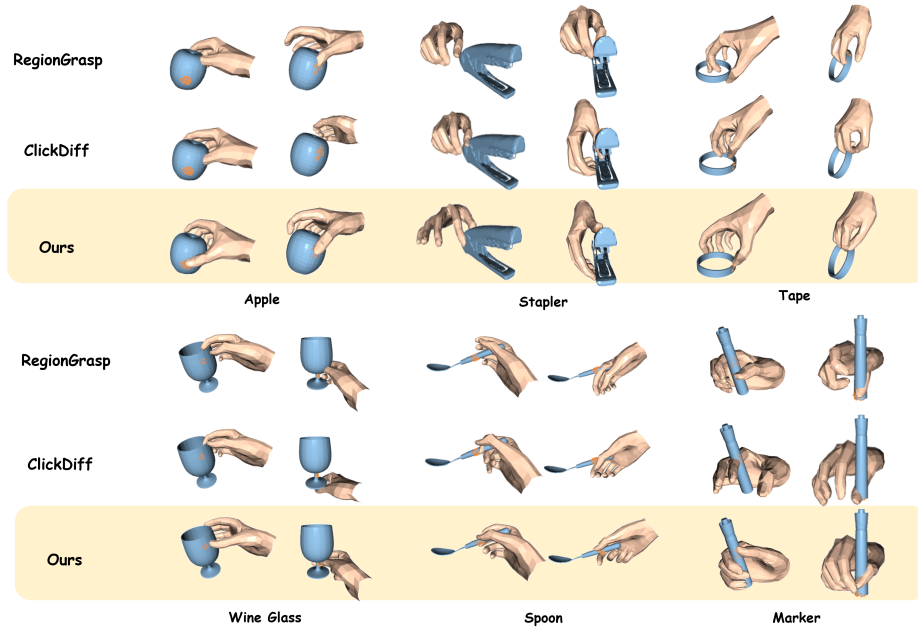


Fig. 3: Qualitative comparison with RegionGrasp [37] and ClickDiff [24]. We select 6 objects, and for each object we visualize two different thumb–contact regions on the object surface (highlighted in orange).

intra-constraint multi-modality, where similar organizations (e.g., Tripod vs. Quadpod) can both satisfy the same thumb-region constraint and remain physically feasible (Fig.5(b)); and (ii) **constraint-induced mode switching on symmetric shapes**, where a small change in the thumb-contact region can reasonably select a different grasp organization (Fig.5(a)). In these cases, even when the Top-1 prediction differs from the single ground-truth label, it can still provide a valid semantic scaffold for downstream synthesis. Consistently, the GT-taxonomy oracle yields only marginal end-to-end gains, as reported in the supplementary material, suggesting limited downstream impact from the remaining Top-1 errors. We nevertheless observe genuine mismatches that lead to clear semantic inconsistency (Fig.5(c)), highlighting remaining challenges under sparse region-level intent inputs.

We empirically validate the importance of taxonomy as a semantic anchor. To isolate the role of **Cognition**, we keep the same architecture but break template consistency by conditioning Stage II/III on a randomly sampled taxonomy label, rather than propagating the predicted one from **Cognition**. This variant leads to a pronounced degradation in generation quality and semantic coherence, and also impairs interaction performance and condition consistency (Tab. 2). Qualitatively, the synthesized grasps exhibit noticeable mode drift, yielding semantically implausible configurations with evident interpenetration (Fig. 6). These results

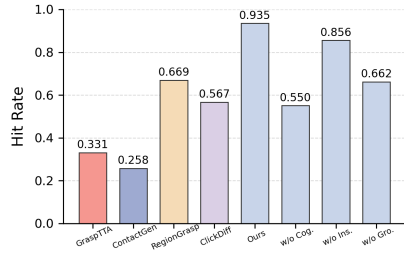


Fig. 4: Condition consistency (Hit Rate, higher is better) for our method, state-of-the-art baselines, and stage-wise ablations.

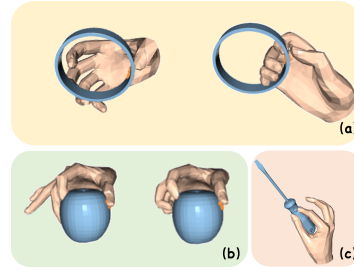


Fig. 5: Taxonomy-classifier reliability under region-level intent. Top-1 deviations are often acceptable due to (a) mode switching on symmetric shapes and (b) intra-constraint multi-modality; (c) shows a true semantic mismatch.

confirm that taxonomy acts as a crucial semantic anchor: without a coherent high-level template, downstream instantiation and grounding cannot reliably preserve grasp semantics or maintain contact-aware geometric alignment.

Stage II: Instantiation. We validate the contribution of diffusion-based pose instantiation by replacing the diffusion-predicted MANO [34] pose parameters with a fixed canonical template pose $\mathbf{p}_{\text{Feix}}$ (defined in the supplementary materials) associated with the inferred Feix taxonomy type, while keeping the rest of the model and the entire pipeline unchanged. Quantitatively, for interaction quality, although penetration-related metrics show a marginal improvement, the contact area decreases and simulation stability deteriorates; meanwhile, for generation quality, both TDD and cFID increase noticeably, indicating a larger discrepancy between the generated results and the real-data distribution (Tab. 2). Qualitatively, the synthesized grasps exhibit pronounced hand-pose distortions (Fig. 6). These findings imply that when the pose initialization is of low quality, the subsequent contact-constrained refinement tends to “reduce contact to alleviate penetration” and the semantic prior encoded by the category template becomes ineffective, degenerating into an optimization driven primarily by contact objectives. This observation substantiates the necessity of diffusion-based instantiation for providing reliable initial poses.

Stage III: Grounding. We validate the importance of the *Grounding* stage by fully ablating it, letting the pipeline proceed directly from *Instantiation* to the final output without geometric refinement. Quantitatively, removing *Grounding* leads to a substantial drop in conditional consistency and interaction quality, along with degraded overall generation quality. Qualitatively, we observe pronounced interpenetration artifacts across multiple regions. These results indicate that, once category-specific contact constraints are enforced, *Grounding* serves as a geometry-aware corrective step that resolves interpenetration and contact

Table 2: Ablation study on the effectiveness of the proposed **Cognition** (Cog.), **Instantiation** (Ins.), and **Grounding** (Gro.) stages.

Cog.	Ins.	Gro.	Interaction				Generation		
			CA $\uparrow$	PD(max) $\downarrow$	PD(mean) $\downarrow$	PV $\downarrow$	SIM $\downarrow$	cFID $\downarrow$	TDD $\downarrow$
$\times$	$\checkmark$	$\checkmark$	<u>116.6</u>	0.364	0.146	1.938	<u>11.34</u>	8.590	1.009
$\checkmark$	$\times$	$\checkmark$	113.2	0.238	0.095	<u>1.461</u>	11.44	3.135	0.444
$\checkmark$	$\checkmark$	$\times$	107.3	0.455	0.197	3.588	11.78	<u>2.625</u>	<u>0.409</u>
$\checkmark$	$\checkmark$	$\checkmark$	120.1	<u>0.340</u>	<u>0.131</u>	1.265	10.72	1.714	0.223

misalignment while preserving pose naturalness and user-specified interaction intent, mitigating the typical realism-geometry trade-off.

4.6 User Study

We conducted a user study to more realistically assess the perceptual performance gap among methods for *grasp pose generation*. We compared **our method** against two representative baselines, **RegionGrasp** [37] and **ClickDiff** [24]. The study recruited 20 participants, including 10 males and 10 females. The Beihang University Institutional Review Board for Biological and Medical Ethics deemed this study exempt under 45 CFR 46.104(d)(3); all participants provided informed consent. To mitigate ordering effects, results from the three methods were presented in a randomized order. In each trial, participants were shown three grasp poses (one generated by each method) under the same thumb-contact region condition, and were asked to rate each result along the following criteria on a 5-point Likert scale [25] (1–5, higher is better): (1) *Interaction quality*, evaluating whether the hand establishes valid contacts for a stable grasp and avoids interpenetration; (2) *Pose quality*, evaluating whether the generated grasp pose is natural, physically plausible, and free from unnatural twisting or deformation; and (3) *Intent consistency*, evaluating whether the grasp pose aligns with human grasping intent under the given thumb-contact condition. For each method, we computed the mean score per criterion by averaging ratings across participants and trials, and report scores obtained by averaging the three criterion scores. As shown in Fig. 7, our method outperforms the other two methods on all three criteria. This indicates that our approach not only generates physically consistent and stable grasps, but also better aligns with human grasp intent under the specified thumb-contact condition. These findings are consistent with the quantitative results.

5 Limitation and Future Work

TaxoGrasp is supervised by HOGraspNet’s Feix taxonomy labels [11], and its thumb-contact regions are inferred post hoc from recorded grasps rather than

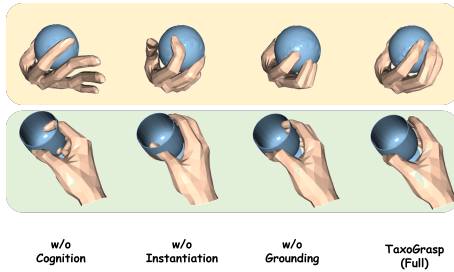


Fig. 6: Qualitative examples from the ablation study on **Cognition**, **Instantiation**, and **Grounding** stages.

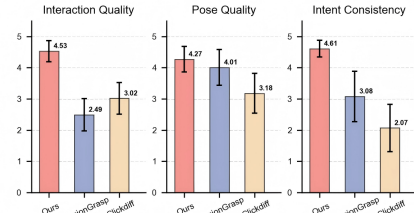


Fig. 7: User study results. The higher score indicates better performance.

collected under controlled thumb-contact specifications. Each training instance therefore contains only the observed taxonomy label for a recorded grasp, without systematic annotation of alternative taxonomy modes compatible with the same object–thumb–region condition. Under this per-instance single-label supervision, our current Top-1 formulation predicts a single taxonomy mode rather than a taxonomy-level multi-modal distribution. Future work will collect a dedicated sparse thumb-contact grasp dataset with repeated grasps under controlled object–thumb–region specifications and explore distributional taxonomy inference to support multi-modal grasp synthesis.

6 Conclusion

We propose **TaxoGrasp**, a novel framework for controllable human grasp synthesis under sparse thumb-contact constraints. By reformulating synthesis as supervised taxonomy inference, our three-stage pipeline integrates taxonomy templates across cognition, instantiation, and grounding to address intent ambiguity and contact-semantics decoupling. This approach yields high-quality grasps that are intent-consistent, physically plausible, and geometrically feasible. Furthermore, we augment HOGraspNet [8] with patch-level thumb-contact annotations and introduce a standardized evaluation protocol. Quantitative and qualitative results demonstrate significant improvements over prior work, validating the effectiveness of our approach.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Project No. 62272019), the China Postdoctoral Science Foundation (Grant No. 2025M774236), the Postdoctoral Fellowship Program of the China Postdoctoral Science Foundation (Grant No. GZC20242159), and the Fundamental Research Funds for the Central Universities, and was also supported by Zhongguancun Laboratory.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
2. Baerlocher, P., Boulic, R.: Parametrization and range of motion of the ball-and-socket joint. *DEFORM/AVATARS* **196**, 180–190 (2000)
3. Balzer, M.: Capacity-constrained voronoi diagrams in continuous spaces. In: 2009 sixth international symposium on Voronoi diagrams. pp. 79–88. IEEE (2009)
4. Cha, J., Kim, J., Yoon, J.S., Baek, S.: Text2hoi: Text-guided 3d motion generation for hand-object interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1577–1585 (2024)
5. Chang, X., Sun, Y.: Text2grasp: Synthesis of grasps by text prompts for object grasping parts. In: International Symposium on Neural Networks. pp. 514–524. Springer (2025)
6. Chen, J., Ke, Y., Peng, L., Wang, H.: Dexonomy: Synthesizing all dexterous grasp types in a grasp taxonomy. arXiv preprint arXiv:2504.18829 (2025)
7. Cheng, S., Jin, Y., Liang, Y., Jiang, L., Wang, H.: An efficient grasping shared control architecture for unpredictable and unspecified tasks. *Frontiers in Neuro-robotics* **18**, 1429952 (2024)
8. Cho, W., Lee, J., Yi, M., Kim, M., Woo, T., Kim, D., Ha, T., Lee, H., Ryu, J.H., Woo, W., et al.: Dense hand-object (ho) graspnet with full grasping taxonomy and dynamics. In: European Conference on Computer Vision. pp. 284–303. Springer (2024)
9. Christen, S., Kocabas, M., Aksan, E., Hwangbo, J., Song, J., Hilliges, O.: D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20577–20586 (2022)
10. Coumans, E., Bai, Y.: PyBullet: A python module for physics simulation for games, robotics and machine learning. <http://pybullet.org> (2021), accessed: 2025-12-30
11. Feix, T., Romero, J., Schmiedmayer, H.B., Dollar, A.M., Kragic, D.: The grasp taxonomy of human grasp types. *IEEE Transactions on human-machine systems* **46**(1), 66–77 (2015)
12. Fernández, U.J., Elizondo, S., Iriarte, N., Morales, R., Ortiz, A., Marichal, S., Ardaiz, O., Marzo, A.: A multi-object grasp technique for placement of objects in virtual reality. *Applied Sciences* **12**(9), 4193 (2022)
13. Fey, M., Lenssen, J.E.: Fast graph representation learning with pytorch geometric. arXiv preprint arXiv:1903.02428 (2019)
14. Grady, P., Tang, C., Twigg, C.D., Vo, M., Brahmabhatt, 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)
15. Han, G., Zhai, W., Yang, Y., Cao, Y., Zha, Z.J.: Touch: Text-guided controllable generation of free-form hand-object interactions. arXiv preprint arXiv:2510.14874 (2025)
16. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
17. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **30**, 6629–6640 (2017)
18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020)

19. Jian, J., Liu, X., Chen, Z., Li, M., Liu, J., Hu, R.: G-dexgrasp: Generalizable dexterous grasping synthesis via part-aware prior retrieval and prior-assisted generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11447–11457 (2025)
20. 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)
21. Li, H., Lin, X., Zhou, Y., Li, X., Huo, Y., Chen, J., Ye, Q.: Contact2grasp: 3d grasp synthesis via hand-object contact constraint. Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence pp. 1053–1061 (2023)
22. Li, H., Nehab, D., Wei, L.Y., Sander, P.V., Fu, C.W.: Fast capacity constrained voronoi tessellation. In: Proceedings of the 2010 ACM SIGGRAPH symposium on Interactive 3D Graphics and Games. pp. 1–1 (2010)
23. Li, K., Wang, J., Yang, L., Lu, C., Dai, B.: Semgrasp: Semantic grasp generation via language aligned discretization. In: European Conference on Computer Vision. pp. 109–127. Springer (2024)
24. Li, P., Wang, Z., Liu, M., Liu, H., Chen, C.: Clickdiff: click to induce semantic contact map for controllable grasp generation with diffusion models. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 273–281 (2024)
25. Likert, R.: A technique for the measurement of attitudes. *Archives of psychology* **22** **140**, 1–55 (1932)
26. Lin, J.: Divergence measures based on the shannon entropy. *Information Theory, IEEE Transactions on* **37**(1), 145–151 (1991)
27. Liu, A.L., Chao, Y.W., Chen, Y.T.: Task-oriented human grasp synthesis via context-and task-aware diffusers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10375–10385 (2025)
28. 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)
29. Ma, Y., Chen, K., ZHENG, K., QI, D.: Contact map transfer with conditional diffusion model for generalizable dexterous grasp generation. In: *Advances in Neural Information Processing Systems*. vol. 38, pp. 79807–79836 (2025)
30. Mandikal, P., Grauman, K.: Dexvip: Learning dexterous grasping with human hand pose priors from video. In: *Conference on Robot Learning*. pp. 651–661. PMLR (2022)
31. Oprea, S., Martinez-Gonzalez, P., Garcia-Garcia, A., Castro-Vargas, J.A., Orts-Escolano, S., Garcia-Rodriguez, J.: A visually realistic grasping system for object manipulation and interaction in virtual reality environments. *Computers & Graphics* **83**, 77–86 (2019)
32. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
33. Ranjan, A., Bolkart, T., Sanyal, S., Black, M.J.: Generating 3d faces using convolutional mesh autoencoders. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 704–720 (2018)
34. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics* **36**(6), 245.1–245.17 (2017)
35. Taheri, O., Choutas, V., Black, M.J., Tzionas, D.: Goal: Generating 4d whole-body motion for hand-object grasping. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 13263–13273 (2022)

36. Wang, P.S., Liu, Y., Tong, X.: Dual octree graph networks for learning adaptive volumetric shape representations. *ACM Transactions on Graphics (TOG)* **41**(4), 1–15 (2022)
37. Wang, Y., Guo, C., Cheng, L., Jiang, H.: Regiongrasp: A novel task for contact region controllable hand grasp generation. In: *European Conference on Computer Vision*. pp. 33–48. Springer (2024)
38. Wang, Y., Zhang, Z., Leng, Z., Liu, H., Li, F.W.B., Li, M., Liang, X.: Multi-modal prior-augmented text-driven 3d human-object interaction generation. *Science China Information Sciences* **69**(5), 1–15 (2026)
39. Wei, Y.L., Jiang, J.J., Xing, C., Tan, X.T., Wu, X.M., Li, H., Cutkosky, M., Zheng, W.S.: Grasp as you say: Language-guided dexterous grasp generation. *Advances in Neural Information Processing Systems* **37**, 46881–46907 (2024)
40. Yang, L., Zhan, X., Li, K., Xu, W., Li, J., Lu, C.: Cpf: Learning a contact potential field to model the hand-object interaction. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 11097–11106 (2021)
41. Zhang, L., Zheng, D., Bai, K., Bing, Z., Marton, Z.C., Chen, Z., Knoll, A.C., Zhang, J.: Omnidexvlg: Learning dexterous grasp generation from vision language model-guided grasp semantics, taxonomy and functional affordance. *arXiv preprint arXiv:2512.03874* (2025)
42. Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: Remodiffuse: Retrieval-augmented motion diffusion model. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 364–373 (2023)
43. Zhang, Z., Wang, H., Yu, Z., Cheng, Y., Yao, A., Chang, H.J.: Nl2contact: Natural language guided 3d hand-object contact modeling with diffusion model. In: *European Conference on Computer Vision*. pp. 284–300. Springer (2024)
44. Zhao, F., Tsetserukou, D., Liu, Q.: Graingrasp: Dexterous grasp generation with fine-grained contact guidance. In: *2024 IEEE international conference on robotics and automation (ICRA)*. pp. 6470–6476. IEEE (2024)
45. Zhao, Z., Qi, M., Ma, H.: Decomposed vector-quantized variational autoencoder for human grasp generation. In: *European Conference on Computer Vision*. pp. 447–463. Springer (2024)
46. Zhou, K., Chen, C., Ma, Y., Leng, Z., Shum, H.P.H., Li, F.W.B., Liang, X.: A mixed reality training system for hand-object interaction in simulated microgravity environments. In: *IEEE International Symposium on Mixed and Augmented Reality*. pp. 167–176. IEEE (2023)
47. Zhou, K., Shum, H.P.H., Li, F.W.B., Liang, X.: Multi-task spatial-temporal graph auto-encoder for hand motion denoising. *IEEE Transactions on Visualization and Computer Graphics* **30**(10), 6754–6769 (2024)