CoToGrasp: Contact-Topology-Conditioned Dexterous Grasp Synthesis via Canonical Workspace Learning

Julien Mérand[1], Boris Meden[1], Liming Chen[2], and Mathieu Grossard[1]

[1] Université Paris-Saclay, CEA, List, F-91120 Palaiseau, France
[2] Ecole Centrale Lyon, CNRS, LIRIS, UMR5205, Institut Universitaire de France (IUF), F-69130 Ecully, France

Abstract. Current dexterous grasp planners primarily optimize for physical stability, focusing on whether an object can be grasped rather than how it should be grasped to support downstream functional tasks. However, conditioning grasp synthesis on specific human grasp taxonomies typically requires prohibitively expensive, object-annotated datasets. To address these limitations, we propose CoToGrasp, a novel generative framework that synthesizes diverse, stable grasps strictly conditioned on specific contact topologies. To bypass the data collection bottleneck, CoToGrasp is trained entirely in an object-agnostic manner. We introduce a feature-based canonical workspace that projects local object features into a unified gripper-centric domain, effectively decoupling the semantic functional intent from the arbitrary object geometry. By learning the intrinsic contact manifold of the gripper within this workspace, our model achieves zero-shot generalization to unseen objects at inference. Extensive evaluations on the large-scale DexGraspNet dataset demonstrate that CoToGrasp achieves state-of-the-art performance, outperforming existing taxonomy-guided planners. Finally, we demonstrate the physical viability and kinematic feasibility of our synthesized contact topologies on a physical robot platform. Code is available on our project website https://cea-list.github.io/cotograspweb/.

1 Introduction

As robotic systems evolve toward embodied agents capable of complex interaction, grasping can no longer be treated solely as a geometric stability optimization problem. The recent rise of humanoid robots, increasingly guided by high-level reasoning frameworks such as Large Language Models (LLMs) [15], demands grasps that directly answer functional, task-level intents. Fine-grained robot manipulation necessitates dexterous, multi-fingered hands with high Degrees of Freedom (DoF) because many downstream tasks – such as in-hand object reorientation, precision tool insertion, finger gaiting, and handle-based grasping – demand controllable, distributed multi-point contact topologies that fundamentally exceed the symmetric pinch and enveloping capabilities of simple parallel-jaw grippers.

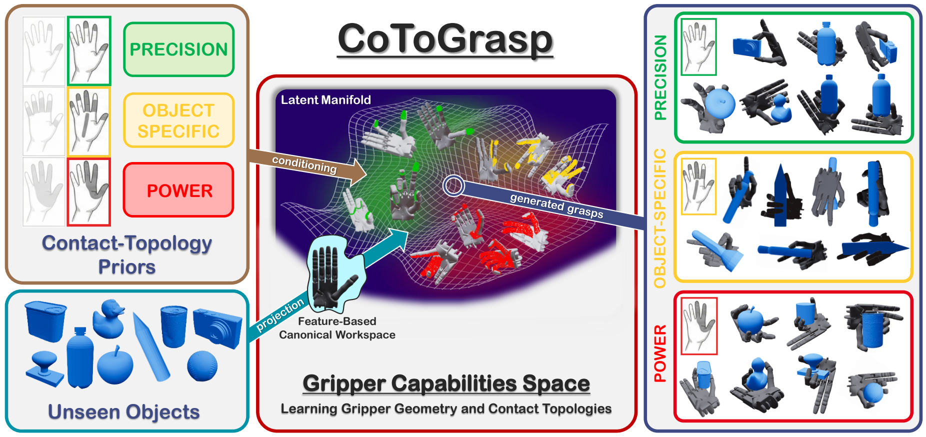

Fig. 1: Contact-Topology-Conditioned Grasp Synthesis. Given a desired semantic contact-topology condition (top left) – categorized into **Precision**, **Object-Specific** (highly constrained topologies tailored for specific tool use) or **Power** functional groups – and a novel, unseen object (bottom left), our framework synthesizes functionally diverse and physically stable grasps (right). Rather than learning contact topologies directly on the object geometry, we project local object features into a feature-based canonical workspace. This unified spatial representation effectively decouples the functional intent from the specific object identity. Within this workspace, we learn a latent manifold (center) that models the intrinsic contact capabilities of the gripper, enabling zero-shot generalization to diverse target geometries.

However, the transition to high-DoF systems exposes a critical limitation in existing grasp planners: while they can produce physically stable grasps, they lack the structural properties required to generate task-aligned contact topologies. Modern data-driven grasp planners predominantly focus on whether an object can be grasped rather than how it should be manipulated. By optimizing purely for geometric stability and force closure, these methods introduce a severe generative bias toward energetically stable but functionally uniform grasps. Although dexterous hands offer rich articulated contact capabilities, current planners do not fully exploit their potential in terms of grasp pattern diversity, instead overwhelmingly defaulting to enveloping power grasps. Consequently, synthesizing a specific precision grasp required for a downstream task becomes highly inefficient, as it necessitates the generation and rejection of an impractical volume of functionally unsuitable candidates.

To achieve task-aligned grasp synthesis, it is necessary to introduce a structural prior over valid contact configurations. In this work, rather than relying on arbitrary stable grasps, we condition grasp synthesis on structured contact topologies derived from human grasp taxonomies. Specifically, we adapt the Gonzalez taxonomy [9], which categorizes grasps based strictly on the hand's active contact surfaces rather than the object's shape. This hand-centric formulation allows for seamless adaptation to anthropomorphic grippers independent of their specific kinematics. Furthermore, to avoid the severe generalization bottlenecks associated with explicitly mapping contact topologies to specific object geometries in training datasets, we propose building on the object-agnostic training paradigm introduced in [28]. By learning the intrinsic contact capabilities of the robotic hand independently of specific objects, we completely decouple grasp semantics from object topology.

We introduce CoToGrasp (Contact-Topology-Conditioned Dexterous Grasp Synthesis via Canonical Workspace Learning), a gripper-oriented framework for generating diverse, contact-topology-conditioned grasps. Unlike standard methods that learn contact maps directly on object point cloud, our approach utilizes a Canonical Feature-based Workspace anchored to the gripper frame. This workspace acts as a domain-agnostic bridge: we first extract local geometric features via a DGCNN [30] encoder from either the gripper (training) or the object (inference), then aggregate these features into fixed points within the canonical workspace using k-Nearest Neighbors (kNN). To capture the complex spatial structure of functional grasps, we treat each populated workspace point as a discrete token and process this set of workspace-anchored embeddings using a Transformer [35] encoder. This attention mechanism allows the network to learn structural constraints by modeling the specific spatial distribution and co-occurrence patterns of contact points inherent to each topology. These refined features are distilled into a compact latent descriptor via a Set Transformer [17], which conditions a Conditional Variational Auto-Encoder (CVAE) [32] to reconstruct probabilistic contact masks. Finally, we synthesize the physical grasp through a test-time energy-based optimization that aligns the gripper's active surface with the predicted contact zones, enforcing kinematic constraints and collision avoidance.

We evaluate our method extensively in both simulation and real-world settings. We first demonstrate the severe functional bias of current taxonomy-unaware planners by measuring the entropy of their retrieved contact topology distributions. We then compare CoToGrasp against state-of-the-art taxonomy-guided methods, demonstrating superior topology compliance and stability.

In summary, our main contribution is the introduction of CoToGrasp, a novel object-agnostic grasp planner that synthesizes grasps conditioned on structured contact topologies derived from human taxonomies. Furthermore, to properly assess these capabilities, we establish a rigorous evaluation methodology to quantify the functional diversity and semantic bias inherent in dexterous grasp planners.

2 Related Work

Learning-Based Grasp Planners. The rise of humanoid robots demands robust, task-oriented multi-fingered grasp synthesis [10, 18]. However, contemporary data-driven grasp planners – whether directly predicting explicit joint configurations [12, 13, 22, 24, 37, 39, 41, 44, 46] or learning intermediate grasp representations [1, 14, 21, 38] – prioritize physical stability over functional intent.

Heavily reliant on synthetic databases [33, 36, 43] generated via analytical force-closure resolution [23, 29], these models frequently suffer from mode collapse. They default to enveloping power grasps, underutilizing hand dexterity. Furthermore, explicitly mapping grippers to specific objects biases these models, hindering shape generalization. To circumvent this, frameworks like GOAG [28] utilize object-agnostic paradigms, demonstrating that learning contact topologies directly within the gripper's canonical space yields superior generalization to novel geometries.

Human Taxonomies and Task-Oriented Semantics. To address how an object should be grasped, researchers draw from biomechanics. Cutkosky [5] established early categorizations based on object properties and task requirements, while Bullock et al. [2] considered the motion dynamics. Feix et al. [8] later proposed the comprehensive GRASP taxonomy. Crucially for haptics, Gonzalez et al. [9] synthesized these frameworks by categorizing 21 distinct contact topologies based strictly on the hand's active contact surfaces.

Concurrently, task-oriented planners have leveraged these taxonomies to inject high-level semantics into robotic actions. Early studies utilized CNNs to decode manipulation semantics [6, 26, 31], while recent approaches employ Vision-Language Models [19,20] to guide grasp selection. However, bridging the gap between high-level language semantics and low-level physical interactions remains challenging. Methods like FunGrasp [11] retarget human-object interactions to grippers, but remain heavily dependent on task-specific interaction priors.

Taxonomy-Conditioned Grasp Synthesis. To explicitly control functional intent, recent methods condition their generative pipelines on taxonomy types. These representations vary from holistic couplings of joints, contacts, and object geometry [16, 25], to manual grouping [40], to purely kinematic "eigen grasps" [7, 27]. Recent frameworks like Dexonomy [4] and OmniDexVLG [45] represent types via explicit joint and contact combinations. Despite these advancements, current planners share a critical limitation: the deeply ingrained coupling of taxonomy types with specific object geometries during dataset construction. Methods like Dexonomy [4] require massive, analytically computed datasets where grasp types are strictly annotated against specific object meshes. If a novel object's local geometry does not perfectly accommodate the rigid precomputed joint template, the optimization fails or collapses into a functionally incorrect type. CoToGrasp subverts this limitation by utilizing a purely hand-centered taxonomy [9] based on semantic contact masks rather than rigid joint configurations. By training our generative model in a strictly object-agnostic manner, we entirely bypass the need for costly, object-annotated datasets.

3 Method

Figure 2 provides an overview of the CoToGrasp framework. Formally, we define a grasp as a tuple $(R, \mathbf{t}, Q)$, where $(R, \mathbf{t}) \in SO(3) \times \mathbb{R}^3$ represents the 6D pose of the gripper relative to the object frame $\mathcal{O}$, and Q denotes the internal joint

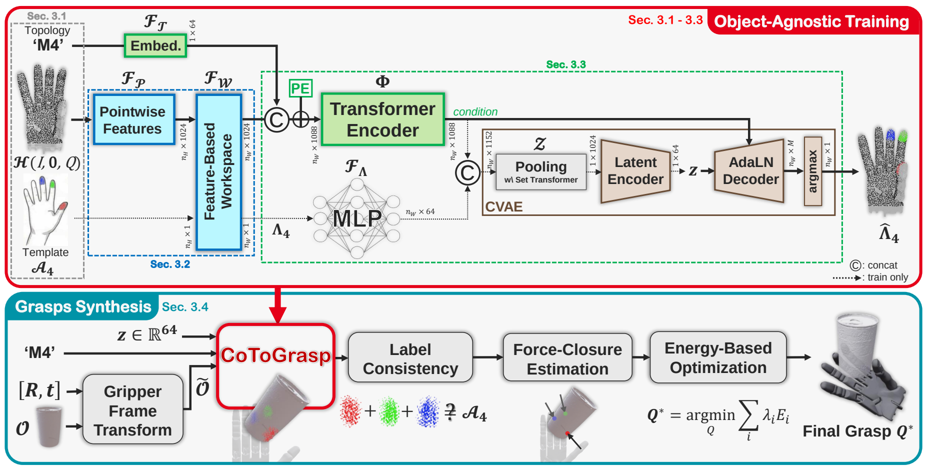

Fig. 2: CoToGrasp Method Overview. The proposed framework operates in two distinct phases. **Top (Object-Agnostic Training):** The model learns an intrinsic, gripper-centric contact manifold within a canonical feature-based workspace, independent of object geometry. **Bottom (Grasp Synthesis):** At inference, a target object is transformed into the canonical frame. The network's contact-topology-conditioned prediction is strictly filtered through a validation pipeline before energy-based optimization aligns the gripper to yield the final stable grasp (Q^*).

configuration. Our primary objective is to synthesize a diverse set of physically stable grasps that strictly respect a requested semantic contact topology.

To succesfully decouple high-level functional semantics from arbitrary object geometries, we adopt an object-agnostic learning paradigm. Unlike standard methods, our architecture is trained exclusively on gripper point clouds and inferred on target object point clouds. During the training phase, the model operates entirely within the canonical gripper frame, rendering it independent of the global pose $(R, \mathbf{t})$. The network takes as input only the gripper's local surface geometry and a semantic contact topology, learning to reconstruct the corresponding physical contact template mask.

3.1 Gripper-Oriented Contact Topology

Let $\mathcal{I} = \{1, \ldots, N_H\}$ be the fixed index set of the points on the gripper's active grasping surface – the specific subset of the gripper geometry designed to establish physical contact – representing a configuration-independent domain. We define the gripper handprint, $\mathcal{H}(R, \mathbf{t}, Q)$, as the spatial embedding of these points given a 6D pose $(R, \mathbf{t})$ and a joint configuration Q:

$$\mathcal{H}(R, \mathbf{t}, Q) = \{\mathbf{h}_i(R, \mathbf{t}, Q) \in \mathbb{R}^6 \mid i \in \mathcal{I}\} \quad (1)$$

where $\mathbf{h}_i(R, \mathbf{t}, Q)$ denotes the global Cartesian coordinates (x, y, z) of the i-th point concatenated with its corresponding surface normal (n_x, n_y, n_z). The hand-

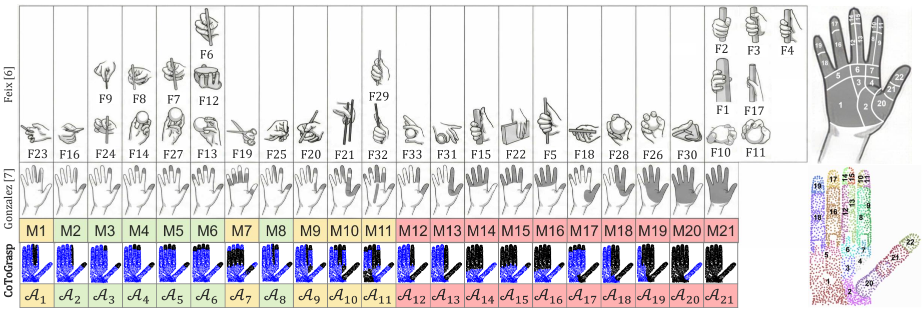

Fig. 3: Semantic Grasp Taxonomy and Contact Mapping. (Left) Correspondences between the classical Feix [8] (F) taxonomy (top), the haptic Gonzalez [9] (M) taxonomy (middle row) and our derived point cloud contact templates $\mathcal{A}_m$ (bottom row). We categorized the 21 templates into three distinct functional groups: **Precision**, **Power** and **Object-Specific** (highly constrained topologies tailored for specific tool use). **(Right)** The 22 anatomical contact zones defined by Gonzalez (top) and the direct surjective mapping (ζ) onto our discrete gripper handprint $\mathcal{H}$ (bottom).

print $\mathcal{H}$ is strictly defined as the discretization of the gripper's active grasping surfaces (detailed in Supp. Mat. A).

We adapt the taxonomy $\mathcal{T}$ from [9], consisting of $M = 21$ pairs of topology names and contact templates. Each template $\mathcal{A}_m : \mathcal{I} \longrightarrow \{0, \ldots, N\}$ acts as a semantic mask, assigning a Zone ID to points required for contact topology m:

$$\mathcal{T} = \{(\text{Name}_m, \mathcal{A}_m) \mid m \in [1, M]\}, \qquad \mathcal{A}_m(i) = \begin{cases} \zeta(i) & \text{if } i \in \mathcal{S}_m \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $\zeta : \mathcal{I} \to \{1, \ldots, N\}$ is a surjective mapping associating every hand point, from $\mathcal{H}$, to a physical zone and $\mathcal{S}_m$ is the set of active points for contact topology m. This formulation, illustrated in Figure 3, enables the systematic adaptation of the human-centric taxonomy [9] to anthropomorphic grippers. Crucially, this taxonomy serves purely as a labeling interface. CoToGrasp easily adapts to any alternative contact-based representation.

3.2 Geometric Transfer via Canonical Workspace Learning

The Duality of Contact. Contact between a gripper and an object is fundamentally a symmetric geometric relation: a point on the gripper surface is in contact with the object if and only if a corresponding object point is in contact with the gripper. [14, 21] formulate contact detection as an object-centric problem, parameterizing the search for valid grasp locations over the object surface $\mathcal{O}$ to define the contact set $\mathcal{C}_{\text{obj}} \subset \mathcal{O}$. Conceptually, they ask, "where on the object can I touch?". However, this duality allows us to invert the paradigm. Instead, we ask, "which regions of my hand are activated by this object?" By flipping the formulation, we define the gripper contact set $\mathcal{C}_{\text{grip}}$ directly on the hand's surface. We reformulate the contact condition as:

$$\mathcal{C}_{\text{grip}}(\mathcal{H}(R,\mathbf{t},Q)) = \{h_i \in \mathcal{H}(R,\mathbf{t},Q) \mid \min_{o\in\mathcal{O}} \mathcal{D}_{\text{aligned}}(h_i,o) < \epsilon\} \tag{3}$$

with the distance $\mathcal{D}_{aligned}$ defined between any two points $\mathbf{x}$ and $\mathbf{y}$, introduced in [21], as:

$$\mathcal{D}_{aligned}(\mathbf{x},\mathbf{y}) = e^{\gamma(1-\langle \mathbf{x}-\mathbf{y}, n_{\mathbf{x}}\rangle)}\sqrt{\|\mathbf{x}-\mathbf{y}\|_2} \tag{4}$$

where $n_{\mathbf{x}}$ the surface normal at $\mathbf{x}$ and γ is a scaling factor modulating the influence of the kinematic alignment compared to the pure Euclidean distance.

This duality matters profoundly because it fundamentally changes the learning domain. The object surface $\mathcal{O}$ represents an unbounded domain with infinite variability and high geometric entropy. In contrast, the gripper surface $\mathcal{H}$ possesses a fixed structure, a bounded domain, and a finite kinematic manifold. By shifting the formulation from an unbounded object space to the fixed, mechanism-intrinsic gripper manifold, the solution space $\mathcal{C}_{\text{grip}}$ is strictly bounded by $\mathcal{H}$, simplifying the generative task. Consequently, the network avoids memorizing arbitrary target geometries, instead learning grasp templates purely as fundamental, object-agnostic capabilities of the gripper.

Shift to a Gripper-Centric Frame. Conceptually, this duality allows us to invert the traditional learning paradigm by anchoring our perspective entirely on a canonical gripper frame, where the palm rests permanently at the origin ($R = I, \mathbf{t} = \mathbf{0}$). Here, the handprint's spatial embedding, $\mathcal{H}(Q)$, depends strictly on the internal joint configuration Q, completely isolated from the global grasp pose. The target object $\mathcal{O}$ is brought into this shared space via the inverse transformation $(R,\mathbf{t})^{-1}$, yielding $\tilde{\mathcal{O}}$. Crucially, because physical contact involves opposing surfaces (Eq. 4), we invert the object's surface normals, transforming its local geometry into a negative mold of the expected gripper surface.

Feature-Based Workspace Representation. A central challenge arises from the fact that training and inference operate on geometrically distinct domains. Directly learning over these heterogeneous domains would entangle contact reasoning with object-specific topology. To resolve this, we introduce a canonical feature-based workspace $\mathcal{W} = \{w_j \in \mathbb{R}^3\}_{j=1}^{N_W}$, defined as a fixed set of spatial basis points anchored to the gripper frame.

To capture fine-grained surface details, we extract a dense local description from the input geometry $\mathcal{P} = \{\mathbf{p}_i \in \mathbb{R}^6\}_{i=1}^{N_P}$. We employ a modified DGCNN [30] to extract point-wise local features $\mathcal{F}_{\mathcal{P}} = \{\mathbf{f}_i\}_{i=1}^{N_P}$.

To disentangle these features from the input topology, we project them onto the fixed basis $\mathcal{W}$ using a weighted k-Nearest Neighbors (kNN) aggregation. To ensure surface orientation strictly dictates contact viability, we gate the aggregation using the aligned distance ($\mathcal{D}_{\text{aligned}}$). Let $\mathcal{N}_j$ denote the indices of the k nearest points in $\mathcal{P}$ to the workspace basis point w_j. The aggregated feature $\mathcal{F}_{\mathcal{W}_j}$ is computed as:

$$\mathcal{F}_{\mathcal{W}_j} = \frac{1}{k}\sum_{i\in\mathcal{N}_j} \mathbf{f}_i \cdot e^{-\mathcal{D}_{\text{aligned}}(\mathbf{p}_i, w_j)} \tag{5}$$

By embedding local geometric features directly into the canonical workspace basis $\mathcal{W}$, we construct a consistent set of spatial tokens, effectively decoupling the semantic contact reasoning from the arbitrary input geometry (further projection details and justifications are provided in Supp. Mat. B).

Projected Ground Truth Templates. To construct the supervisory signal used during training, we project the templates $\mathcal{A}_m$ onto the canonical workspace $\mathcal{W}$. We strictly gate this projection using the distance threshold $\epsilon = 0.8$, consistent with our contact formulation in Eq. 3. Each workspace point w_j is assigned a label according to its nearest kinematically aligned hand point:

$$\Lambda_m(j) = \begin{cases} \mathcal{A}_m(i^*) & \text{if } \mathcal{D}_{\text{aligned}}(\mathbf{h}_{i^*}(Q), w_j) < \epsilon \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where $i^* = \underset{i \in \mathcal{I}}{\operatorname{argmin}}\ \mathcal{D}_{\text{aligned}}(\mathbf{h}_i(Q), w_j)$.

Workspace points lying beyond this threshold are assigned the label 0, indicating no physical contact. Consequently, $\Lambda_m \in \{0, \ldots, N\}^{N_W}$ serves as the explicit ground truth contact map.

3.3 Learning Contact Topology Templates using Self-Attention

To model non-local dependencies between potential contact regions, we process the workspace embeddings using a Transformer [35] encoder. To prevent the semantic signal of the requested contact topology from diluting across deep attention layers, we explicitly concatenate the learnable topology embedding $\mathcal{F}_\mathcal{T}$ to every workspace point feature $\mathcal{F}_\mathcal{W}$, denoted $\mathcal{F}_{\mathcal{W} \oplus \mathcal{T}}$. We inject spatial awareness via sinusoidal positional encodings (PE), allowing the multi-head attention mechanism to recover spatial relationships purely from the fixed basis indices:

$$\Phi = \text{Transformer}(\mathcal{F}_{\mathcal{W} \oplus \mathcal{T}} + \text{PE}) \tag{7}$$

To bridge the modality gap, the projected templates Λ_m are passed through a Multi-Layer Perceptron (MLP) to create a continuous embedding $\mathcal{F}_\Lambda$. These are fused with the transformer features Φ and compressed into a global latent descriptor $\mathcal{Z}$ using a Set Transformer [17]. We leverage a Pooling by MultiHead Attention (PMA) block and a Set Attention Block (SAB) to capture complex structural dependencies:

$$\mathcal{Z} = \text{SAB}(\text{PMA}_s(\Phi \oplus \mathcal{F}_\Lambda)) \tag{8}$$

Finally, a CVAE [32] models the localized, intra-topology stochasticity of the contact patches. Because macroscopic grasp multi-modality is explicitly resolved by the topology conditioning, a standard unimodal Gaussian prior $\mathcal{N}(0, I)$ proves highly efficient. A convolutional encoder maps $\mathcal{Z}$ to the posterior distribution $q(z|\mathcal{Z})$, from which a latent variable $z \in \mathbb{R}^\psi$ is sampled. The decoder, conditioned on Φ via Adaptive Normalization [42] layers (AdaLN), modulates the features to output the predicted zone probabilities $\hat{\Lambda}_m$.

The network is trained end-to-end to minimize the standard variational lower bound (ELBO), consisting of a Multi-Class Cross-Entropy reconstruction loss over the workspace points and a Kullback-Leibler regularization term weighted by β. (See Supp. Mat. B for architectural justifications and loss formulations.)

3.4 Grasp Synthesis Through Contact Points Generation

In the inference phase, our goal is to synthesize a valid joint configuration Q that realizes a specific contact topology m for an object $\mathcal{O}$ at a candidate pose $(R, \mathbf{t})$. We first transform the object point cloud into the canonical gripper frame via the inverse rigid transformation $(R, \mathbf{t})^{-1}$. This transformed geometry $\tilde{\mathcal{O}}$, along with the requested contact topology m, is fed into the network – bypassing the latent encoder by sampling z from the prior $\mathcal{N}(0, I)$ – to predict a contact template $\hat{\Lambda}_m$ over the workspace. To ensure the physical viability of the grasp before costly optimization, we enforce a strict cascading validation pipeline:
Label-Consistency Check. A contact topology is fundamentally constrained by the object's local geometry. To prevent generating ill-posed grasps, we apply a template-consistency filter immediately after inference. We perform a strict binary check on the symmetric difference between the ground-truth (Λ_m) and predicted ($\hat{\Lambda}_m$) templates. We allow a maximum deviation of one missing or hallucinated contact zone, safely rejecting fundamentally distinct failures where the object cannot support the requisite contacts.
Contact Points Force-Closure Validation. For candidates passing the label-consistency check, we assess the potential grasp stability using the rapid, approximated force-closure estimation proposed in [28]. Specifically, we extract the 3D workspace points $w_j \in \mathcal{W}$ assigned a non-zero label by $\hat{\Lambda}_m$. We compute the barycenters of these active regions to construct the theoretical grasp wrench space. Crucially, if the force-closure condition is not met, it implies the proposed contact distribution is physically unviable. In this case, we discard the prediction and exploit the generative capacity of our network by resampling the latent variable z to produce a fresh contact template $\hat{\Lambda}_m$. We limit this resampling process to a maximum of 20 iterations to prevent exhaustive searches in poorly conditioned poses. (Detailed formulations are provided in Supp. Mat. D).
Joint Configuration Optimization. Finally, we retrieve the optimal joint configuration Q^* by minimizing an energy function that aligns the gripper's active surfaces with the predicted 3D spatial workspace points, while respecting kinematic constraints:

$$Q^* = \underset{Q}{\operatorname{argmin}} \left(\lambda_d E_{dist} + \lambda_p E_{pen} + \lambda_s E_{spen} + \lambda_j E_j + \lambda_r E_{rep}\right) \tag{9}$$

We adopt the standard terms E_{dist} (contact alignment), E_{pen} (object penetration), E_{spen} (self-collision), and E_j (joint limits) used in [14,21,28]. Additionally, to best match the desired contact topology without prior shape knowledge, we introduce a repulsive term E_{rep}. This term creates a repulsive field around the object for all unused fingers and palm regions (i.e., those not assigned a valid label in $\hat{\Lambda_m}$), enforcing a minimum safety margin of 5 mm to guide the optimizer toward strictly topology-compliant, collision-free configurations. (Detailed mathematical formulations and weighting factors are provided Supp. Mat. D).

4 Experiments

4.1 Experimental Setup

Training Data Formulation. To construct the object-agnostic training dataset, we first define the semantic mapping between the physical surface regions of the target gripper and the required active contact zones ($N = 22$) for the $M = 21$ contact topology templates. We uniformly sample $10,000$ kinematically valid joint configurations Q and compute their corresponding spatial handprints $\mathcal{H}(Q)$ via forward kinematics. By pairing every sampled handprint with all 21 contact templates, we construct a comprehensive training dataset of 210,000 unique data points. Crucially, because this formulation operates entirely within the canonical gripper frame, dataset generation requires zero object meshes or computationally expensive physical simulations. (Further implementation and training details are provided in Supp. Mat. A).

Grasp Pose Constraints and Generation. Because CoToGrasp operates in a canonical, gripper-centric workspace (Sec. 3.2), it inherently decouples the global 6D grasp pose $(R, \mathbf{t}) \in SO(3) \times \mathbb{R}^3$ from local contact generation. By treating this pose as an external prior rather than inferring it end-to-end, our architecture allows task planners or human operators to dictate functionally suitable approaches. This modularity facilitates advanced applications like kinematic keyframing for in-hand manipulation, enabling the synthesis of continuous, topology-consistent grasps along defined object trajectories.

To autonomously evaluate CoToGrasp without external planners, we sample diverse candidate poses using a topology-conditioned heuristic depending on whether the requested topology involves the palm or relies strictly on distal precision. (Detailed sampling formulations are provided in Supp. Mat. E).

Automatic Topology Recognition from Grasp. To verify that CoToGrasp accurately synthesizes the specified contact topologies, we introduce an automated classification step for the generated grasps. First, we extract the subset of gripper points in physical contact with the object (using Eq. 3) and map them back to their semantic zone identifiers (visible in Fig. 3). We then compare these observed active zones against the ground-truth required zones of each target template $\mathcal{A}_m$. We evaluate this match by computing a similarity score s_m based on the Tversky index [34]. Crucially, we apply an asymmetric penalization that punishes extra, unintended contact regions more strictly than missing ones, effectively penalizing clumsy or degenerated grasps. Finally, a generated grasp is evaluated against all M templates and assigned the class m yielding the highest similarity score. To ensure robustness and mitigate false positives, any grasp failing to reach a minimum similarity threshold ($s_m < 0.5, \forall m \in [1, M]$) is strictly classified as "*unknown*". (Detailed mathematical formulations of this metric are provided in Supp. Mat. F).

Method	Object-Agnostic Training	SR ↑	$\mathbf{H}_{TC}$ ↑	Speed (sec./grasps)	Diversity (avg.) ↑		
					R (rad)	Q (rad)	
DFC [23]	✓	72.15	0.7389	>1800	0.0607	1.424	**0.3579**
GenDexGrasp [21]	✗	71.15	0.5956	14.65	0.0519	1.416	0.2567
DRO-Grasp [38]	✗	63.30	0.6504	1.72	0.0546	**1.515**	0.2892
GOAG [28]	✓	**77.90**	0.6527	0.20	0.0479	1.401	0.3170
CoToGrasp	✓	36.94	**0.83**	**0.11**	**0.0674**	1.4927	0.3458

Table 1: Comparison with taxonomy-unaware baselines. CoToGrasp achieves the highest semantic entropy ($\mathbf{H}_{TC}$) and generation speed, overcoming the functional mode collapse typical of unconditioned planners.

4.2 Evaluation Metrics

To evaluate CoToGrasp, we utilize a combination of standard physical stability metrics and novel semantic compliance metrics. For physical stability (Success Rate, **SR**), Generation **Speed**, and Spatial **Diversity**, we strictly follow the evaluation protocols established in [21, 28, 38] (Detailed formulations for these metrics are provided in Supp. Mat. G).

To quantify the functional accuracy and distributional fairness of the generated grasps, we introduce the following novel metrics:

Topology Compliance (TC). This metric evaluates how accurately the generated grasp adheres to the functional constraints of the requested contact topology. Across the M evaluated topologies, the average **TC** is defined as:

$$\mathbf{TC} = \frac{1}{M} \sum_{m=1}^{M} \frac{\text{Effective}_m}{\text{Attempt}_m} \tag{10}$$

where Attempt_m is the total number of simulated grasps conditioned on target topology m, and Effective_m is the subset of those grasps that successfully satisfy the valid semantic contact template for topology m.

Entropy. To evaluate the topological diversity and assess potential mode collapse, we compute the normalized Shannon entropy across all contact topologies:

$$\mathbf{H} = \frac{-\sum_{m=1}^{M} \tilde{f}_m \ln(\tilde{f}_m)}{\ln(M)} \tag{11}$$

A value of $\mathbf{H} \to 1$ indicates a uniform, unbiased generative distribution, whereas $\mathbf{H} \to 0$ reflects severe mode collapse toward a few dominant topologies. We report two distinct variants of this metric:

Stability Entropy ($\mathbf{H}_{SR}$): Evaluates if the model generates physically stable grasps equally well across all topologies. Here, $\tilde{f}_m$ is the normalized **SR** of topology m, defined as $\tilde{f}_m = \mathbf{SR}_m / \sum_{j=1}^{M} \mathbf{SR}_j$.

Semantic Entropy ($\mathbf{H}_{TC}$): Evaluates if the model preserves true functional intent without defaulting to simpler power grasps. Here, $\tilde{f}_m$ is the normalized **TC** of the topology m, defined as $\tilde{f}_m = \mathbf{TC}_m / \sum_{j=1}^{M} \mathbf{TC}_j$.

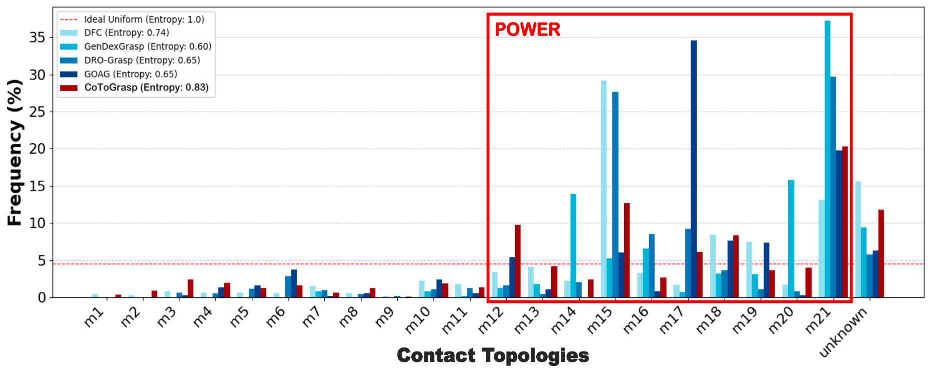

Fig. 4: Functional contact topology distribution across taxonomy-unaware planners. The histogram illustrates the distribution of grasps generated by unconditioned baselines compared to CoToGrasp on the Multidex objects set. The unknown category represents physically stable grasps with unnatural contact patterns that fail to match any contact topology. Notably, unconditioned baselines exhibit a severe generative bias (mode collapse) toward enveloping power grasps (red box).

4.3 Overall Performances

We evaluate CoToGrasp against state-of-the-art baselines using the Shadow Hand. First, we evaluate against unconditioned generative planners to highlight the necessity of topology conditioning. Second, we evaluate against a taxonomy-aware baseline to demonstrate CoToGrasp's superior functional control. For all experiments, we cap generation at 20 attempts per topology. This resampling budget mitigates heuristic approach poses $(R, \mathbf{t})$ that are geometrically incompatible with the requested contact topology.

Taxonomy-Unaware Grasp Planners Analysis and Comparison. Evaluating on a MultiDex subset, Table 1 shows unconditioned baselines [21,23,28,38] achieve higher success rates (SR) by naturally defaulting to geometrically stable, enveloping power grasps (Fig. 4). Consequently, they struggle with precision grasps, yielding low semantic entropy ($\mathbf{H}_{TC}$). Conversely, CoToGrasp achieves a balanced topology distribution and the highest $\mathbf{H}_{TC}$. Its overall SR (36.94%, with a Top-5 average of 56.18%) occurs because it is forced to synthesize diverse topologies (e.g., precision pinches) regardless of an object's natural geometric affordances, inherently prioritizing target contacts over simulated physics stability. While methods like DFC show high spatial variance $(\mathbf{t}, R, Q)$, this primarily reflects joint range exploitation rather than functional diversity. Furthermore, CoToGrasp demonstrates the fastest generation speed at 0.11s per grasp.

Taxonomy-Aware Grasp Synthesis. To evaluate precise functional control, we compare CoToGrasp against Dexonomy [4] on the DexGraspNet [36] test set. After mapping Dexonomy's Feix [8] taxonomy to our contact-based representations for direct comparison (Fig. 3), Table 2 shows CoToGrasp achieves higher **SR** and **TC** (17.18% vs. 14.28%) across all functional categories. While

Method	SR (%)					$\mathbf{H}_{SR}$	TC (%)	$\mathbf{H}_{TC}$
	Power	Precision	Obj. Spe.	Avg. Topo.	Avg. Obj.			
Dexonomy [4]	27.16	12.36	19.62	21.13	23.80	0.91	14.28	0.77
CoToGrasp	**29.75**	**22.71**	**25.50**	**26.72**	**27.56**	**0.96**	**17.18**	**0.84**
CoToGrasp (w/o Label-Consistency)	25.11	14.77	20.85	21.14	22.97	0.94	14.45	0.81
CoToGrasp (w/o Force-Closure)	26.90	14.87	21.08	22.08	23.65	0.95	16.26	0.81
CoToGrasp (No Check)	25.09	14.73	20.58	21.06	23.00	0.94	14.72	0.81

Table 2: Taxonomy-aware grasp synthesis. CoToGrasp outperforms the baseline in both physical stability (SR) and semantic accuracy (TC), particularly on highly constrained precision grasps.

absolute **TC** scores appear modest, they reflect the extreme stringency of our asymmetric metric, which heavily penalizes the minor, physically necessary finger adjustments naturally occurring during dynamic simulation. Outperforming the baseline under these strict conditions demonstrates CoToGrasp's superior zero-shot functional control. Dexonomy [4] struggles with **TC** because its conditioning is rigidly coupled to explicit joint configurations end-to-end; when object geometry forces deviations, the resulting grasp frequently violates the requested topology. Conversely, CoToGrasp robustly projects valid semantic contact zones via object-agnostic training. The ablation of our validation pipeline (No Check) demonstrates a sharp performance drop, emphasizing that verifying topological viability against the object's local geometry via the Label-Consistency check is critical before optimization. (A detailed analysis of **TC** bottlenecks before and after physics simulation is provided in Supp. Mat. I).

Method	SR (%)					
	M2	**M6**	**M11**	**M13**	**M18**	**M21**
Dexonomy [4]	10.5	15.2	60.5†	20.3	29.6	37.2†
CoToGrasp	**30.3**	**21.7**	29.8	**29.6**	**31.3**	33.5

Table 3: Per-Topology results: 3 Power grasps (M13, M18, M21), 2 Precision (M2, M6) and 1 Object-Specific (M11). †Indicates artificially inflated scores due to mode collapse, where Dexonomy defaults to unverified enveloping grasps.

Per-Topology Analysis. Table 3 highlights CoToGrasp's distinct advantage in synthesizing highly constrained precision grasps (e.g. M2: 30.3% vs. 10.5%). An apparent pseudo-anomaly occurs with topologies M11 (Object-Specific) and M21 (Power), where Dexonomy [4] reports artificially inflated **SR** (60.5% and 37.2%). Qualitative analysis reveals this is driven by severe mode collapse: when faced with challenging geometries, Dexonomy [4] frequently abandons the conditional query and defaults to an unverified M21 enveloping grasp. Because it lacks a strict posterior topological check, it erroneously records these power grasps as successes for different input topologies (like M11). Conversely, CoToGrasp explicitly detects and rejects hallucinated contacts, maintaining a competitive true-positive rate for M21 (33.5%) without sacrificing the integrity of the true **TC** metric on precision tasks.

4.4 Real-World Validation

To demonstrate the physical viability and kinematic feasibility of the grasps synthesized by CoToGrasp, we conducted real-world validation experiments. These

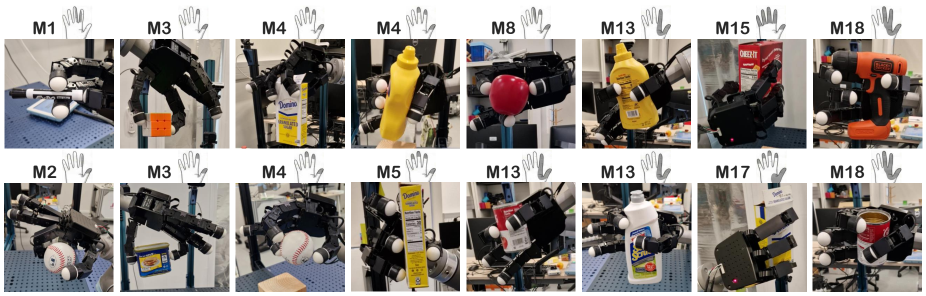

Fig. 5: Real-World kinematic validation. CoToGrasp synthesizes diverse, topology-compliant grasps that are physically executable on a physical Allegro Hand using YCB [3] objects. The target contact topologies (indicated above each frame) demonstrate the physical viability of the generated grasps across both precision and power categories.

were performed in a tabletop environment using a 6-DoF Universal Robots UR10 manipulator equipped with an Allegro Right Hand. Because the Allegro Hand is a four-fingered anthropomorphic gripper (lacking a little finger), we explicitly suppressed the generation of contact topology M6, as it strictly requires five fingers to form a valid contact template. We successfully planned and executed grasps on diverse objects from the YCB [3] dataset. Crucially, we empirically validated the functional diversity of our method by demonstrating that the generated grasp for various contact topologies can be successfully formed and held statically on the physical system. A qualitative overview of the successful real-world grasps, categorized by their respective contact topologies, is presented in Figure 5. Execution sequences are provided in the supplementary video material. The inherent hardware challenges and control limitations of executing precise semantic grasps using a multifingered gripper like the Allegro Hand are discussed extensively in Supp. Mat. J.

5 Conclusion

This paper introduces CoToGrasp, a novel generative framework for contact-topology-conditioned dexterous grasp synthesis. By fundamentally decoupling functional intent from object identity, our approach effectively mitigates the severe mode collapse observed in contemporary grasp planners. This allows CoToGrasp to synthesize highly constrained, topology-compliant precision and power grasps on unseen geometries without relying on costly object-annotated datasets. Extensive evaluations demonstrate that our validation pipeline robustly filters topologically invalid configurations, yielding state-of-the-art semantic diversity and physical stability. Finally, successful real-world deployments on the Allegro Hand confirm that the structural properties of our generated contact topologies are physically executable on a real robot platform.

Acknowledgments

This publication was made possible by the use of the FactoryIA supercomputer, financially supported by the Ile-De-France Regional Council. Experiments presented in this paper were carried out thanks to a platform funded by DIM AI4IDF and PRAIRIE-PSAI. The authors would like to thank Timothée Carecchio for his valuable assistance in achieving the experimental results presented in this paper. This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement nº 101135708 (JARVIS Project). Liming Chen in this research was in part supported by the French Research Agency ANR, l'Agence Nationale de Recherche, through the projects Aristotle (ANR-21-FAI1-0009-01), Astérix (ANR-23-EDIA-0002), DEMETER (ANR-25-HTCE-0002) and PROTEUS (ANR-25-TSIA-0011-01), and the French national investment prioritary program through the PSPC FAIR WASTE project.

References

1. Attarian, M., Asif, M.A., Liu, J., Hari, R., Garg, A., Gilitschenski, I., Tompson, J.: Geometry matching for multi-embodiment grasping. In: Conference on Robot Learning (2023)
2. Bullock, I.M., Ma, R.R., Dollar, A.M.: A hand-centric classification of human and robot dexterous manipulation. IEEE transactions on Haptics (2012)
3. Calli, B., Singh, A., Walsman, A., Srinivasa, S., Abbeel, P., Dollar, A.M.: The ycb object and model set: Towards common benchmarks for manipulation research. In: 2015 international conference on advanced robotics (ICAR) (2015)
4. Chen, J., Ke, Y., Peng, L., Wang, H.: Dexonomy: Synthesizing all dexterous grasp types in a grasp taxonomy. Robotics: Science and Systems (2025)
5. Cutkosky, M.R., et al.: On grasp choice, grasp models, and the design of hands for manufacturing tasks. IEEE Transactions on robotics and automation (1989)
6. Deng, Z., Fang, B., He, B., Zhang, J.: An adaptive planning framework for dexterous robotic grasping with grasp type detection. Robotics and Autonomous Systems (2021)
7. Fang, H.S., Yan, H., Tang, Z., Fang, H., Wang, C., Lu, C.: Anydexgrasp: General dexterous grasping for different hands with human-level learning efficiency. arXiv preprint arXiv:2502.16420 (2025)
8. 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 (2015)
9. Gonzalez, F., Gosselin, F., Bachta, W.: Analysis of hand contact areas and interaction capabilities during manipulation and exploration. IEEE transactions on haptics (2014)
10. Gu, Z., Li, J., Shen, W., Yu, W., Xie, Z., McCrory, S., Cheng, X., Shamsah, A., Griffin, R., Liu, C.K., et al.: Humanoid locomotion and manipulation: Current progress and challenges in control, planning, and learning. IEEE/ASME Transactions on Mechatronics (2026)
11. Huang, L., Zhang, H., Wu, Z., Christen, S., Song, J.: Fungrasp: Functional grasping for diverse dexterous hands. IEEE Robotics and Automation Letters (2025)

12. Huang, S., Wang, Z., Li, P., Jia, B., Liu, T., Zhu, Y., Liang, W., Zhu, S.C.: Diffusion-based generation, optimization, and planning in 3d scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
13. 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 (2021)
14. Khargonkar, N., Casas, L.F., Prabhakaran, B., Xiang, Y.: Robotfingerprint: Unified gripper coordinate space for multi-gripper grasp synthesis and transfer. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2025)
15. Kim, Y., Kim, D., Choi, J., Park, J., Oh, N., Park, D.: A survey on integration of large language models with intelligent robots. Intelligent Service Robotics (2024)
16. Kleer, N., Keil, O., Feick, M., Gomaa, A., Schwartz, T., Feld, M.: Incorporation of the intended task into a vision-based grasp type predictor for multi-fingered robotic grasping. In: 2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN) (2024)
17. Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., Teh, Y.W.: Set transformer: A framework for attention-based permutation-invariant neural networks. In: International conference on machine learning, PMLR (2019)
18. Li, G., Wang, R., Xu, P., Ye, Q., Chen, J.: The developments and challenges towards dexterous and embodied robotic manipulation: A survey. IEEE Robotics & Automation Magazine (2025)
19. Li, H., Mao, W., Deng, W., Meng, C., Fan, H., Wang, T., Osamu, Y., Tan, P., Wang, H., Deng, X.: Multi-graspllm: A multimodal llm for multi-hand semantic guided grasp generation. arXiv preprint arXiv:2412.08468 (2024)
20. Li, K., Wang, J., Yang, L., Lu, C., Dai, B.: Semgrasp: Semantic grasp generation via language aligned discretization. In: European Conference on Computer Vision (2024)
21. Li, P., Liu, T., Li, Y., Geng, Y., Zhu, Y., Yang, Y., Huang, S.: Gendexgrasp: Generalizable dexterous grasping. In: 2023 IEEE International Conference on Robotics and Automation (ICRA) (2023)
22. Liu, M., Pan, Z., Xu, K., Ganguly, K., Manocha, D.: Deep differentiable grasp planner for high-dof grippers. arXiv preprint arXiv:2002.01530 (2020)
23. Liu, T., Liu, Z., Jiao, Z., Zhu, Y., Zhu, S.C.: Synthesizing diverse and physically stable grasps with arbitrary hand structures using differentiable force closure estimator. IEEE Robotics and Automation Letters (2021)
24. Lu, J., Kang, H., Li, H., Liu, B., Yang, Y., Huang, Q., Hua, G.: Ugg: Unified generative grasping. In: European Conference on Computer Vision (2024)
25. Lu, Q., Hermans, T.: Modeling grasp type improves learning-based grasp planning. IEEE Robotics and Automation Letters (2019)
26. Lu, Q., Van der Merwe, M., Sundaralingam, B., Hermans, T.: Multifingered grasp planning via inference in deep neural networks: Outperforming sampling by learning differentiable models. IEEE Robotics & Automation Magazine (2020)
27. Mao, C., Yuan, H., Huang, Z., Xu, C., Ma, K., Lu, Z.: Demofungrasp: Universal dexterous functional grasping via demonstration-editing reinforcement learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 986–995 (2026)
28. Mérand, J., Meden, B., Grossard, M., Chen, L.: Goag: Generative and object-agnostic grasp planner for dexterous robotic manipulation. In: IEEE Int. Conf. Intelligent Robots and Systems (2026)

29. Miller, A.T., Allen, P.K.: Graspit! a versatile simulator for robotic grasping. IEEE Robotics & Automation Magazine (2004)
30. Phan, A.V., Le Nguyen, M., Nguyen, Y.L.H., Bui, L.T.: Dgcnn: A convolutional neural network over large-scale labeled graphs. Neural Networks (2018)
31. Rao, A.B., Krishnan, K., He, H.: Learning robotic grasping strategy based on natural-language object descriptions. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2018)
32. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. Advances in neural information processing systems (2015)
33. Turpin, D., Zhong, T., Zhang, S., Zhu, G., Heiden, E., Macklin, M., Tsogkas, S., Dickinson, S.J., Garg, A.: Fast-grasp'd: Dexterous multi-finger grasp generation through differentiable simulation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA) (2023)
34. Tversky, A.: Features of similarity. Psychological review, American Psychological Association (1977)
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems (2017)
36. Wang, R., Zhang, J., Chen, J., Xu, Y., Li, P., Liu, T., Wang, H.: Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA) (2023)
37. 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 (2024)
38. Wei, Z., Xu, Z., Guo, J., Hou, Y., Gao, C., Cai, Z., Luo, J., Shao, L.: $\mathcal{D}(\mathcal{R},\mathcal{O})$ grasp: A unified representation of robot and object interaction for cross-embodiment dexterous grasping. In: 2025 IEEE International Conference on Robotics and Automation (ICRA) (2025)
39. Weng, Z., Lu, H., Kragic, D., Lundell, J.: Dexdiffuser: Generating dexterous grasps with diffusion models. IEEE Robotics and Automation Letters (2024)
40. Wu, R., Zhu, T., Lin, X., Sun, Y.: Cross-category functional grasp transfer. IEEE Robotics and Automation Letters (2024)
41. Xu, G.H., Wei, Y.L., Zheng, D., Wu, X.M., Zheng, W.S.: Dexterous grasp transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
42. Xu, J., Sun, X., Zhang, Z., Zhao, G., Lin, J.: Understanding and improving layer normalization. Advances in neural information processing systems (2019)
43. Xu, Y., Wan, W., Zhang, J., Liu, H., Shan, Z., Shen, H., Wang, R., Geng, H., Weng, Y., Chen, J., 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 (2023)
44. Xu, Z., Gao, C., Liu, Z., Yang, G., Tie, C., Zheng, H., Zhou, H., Peng, W., Wang, D., Hu, T., et al.: Manifoundation model for general-purpose robotic manipulation of contact synthesis with arbitrary objects and robots. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2024)
45. 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)

46. Zhong, Y., Jiang, Q., Yu, J., Ma, Y.: Dexgrasp anything: Towards universal robotic dexterous grasping with physics awareness. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2025)