arXiv:2608.19776v2 [cs.RO] 21 Aug 2026

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) [18], 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.

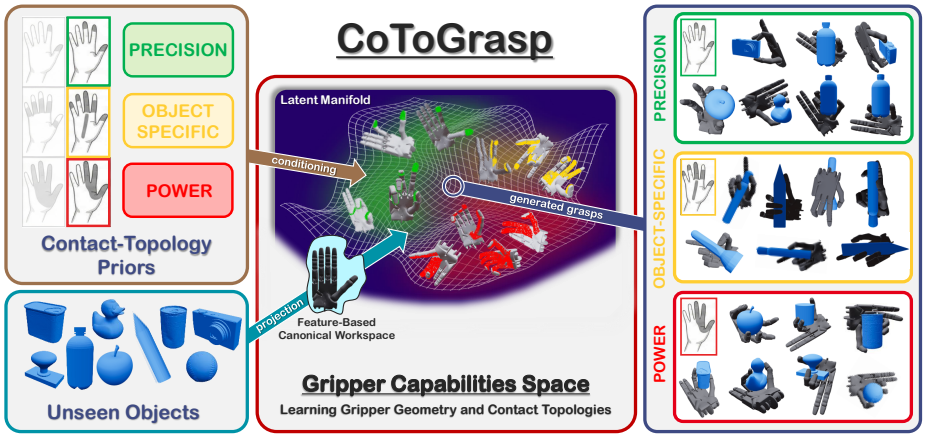

Fig. 1: Contact-Topology-Conditioned Grasp Synthesis. Given a desired semantic contact-topology condition (top left) – categorized into **Precision**, **Object-Specific** (highly constrained topologies tailored for specific tool use) or **Power** functional groups – and a novel, unseen object (bottom 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 [12], 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 [32]. 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 [34] 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 [40] 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 [20], which conditions a Conditional Variational Auto-Encoder (CVAE) [37] 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 [13, 21]. However, contemporary data-driven grasp planners – whether directly predicting explicit joint configurations [15, 16, 25, 27, 42, 44, 46, 50, 52] or learning intermediate grasp representations [1, 17, 24, 43] – prioritize physical stability over functional intent.

Heavily reliant on synthetic databases [38, 41, 49] generated via analytical force-closure resolution [26, 33], 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 [32] 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 [7] established early categorizations based on object properties and task requirements, while Bullock et al. [3] considered the motion dynamics. Feix et al. [11] later proposed the comprehensive GRASP taxonomy. Crucially for haptics, Gonzalez et al. [12] 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 [8, 29, 36], while recent approaches employ Vision-Language Models [22, 23] to guide grasp selection. However, bridging the gap between high-level language semantics and low-level physical interactions remains challenging. Methods like FunGrasp [14] 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 [19, 28], to manual grouping [45], to purely kinematic "eigen grasps" [10, 31]. Recent frameworks like Dexonomy [5] and OmniDexVLG [51] 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 [5] 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 [12] 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

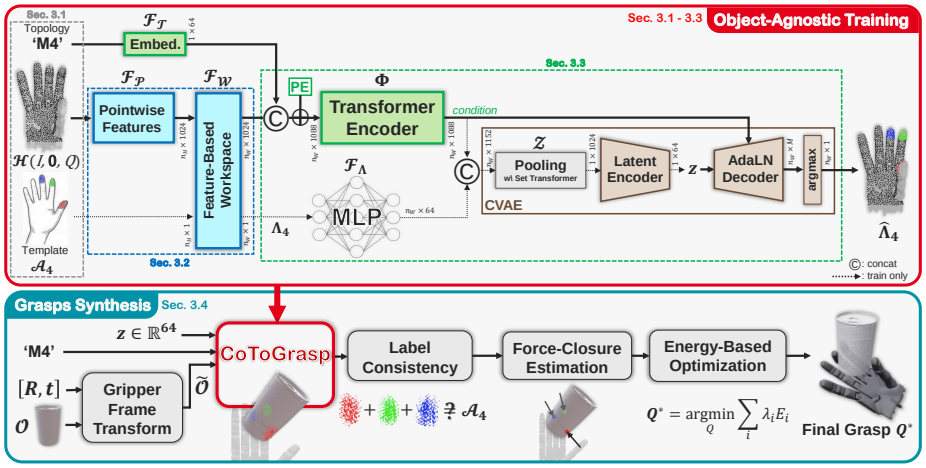

Fig. 2: CoToGrasp Method Overview. The proposed framework operates in two distinct phases. **Top (Object-Agnostic Training):** The model learns an intrinsic, gripper-centric contact manifold within a canonical feature-based workspace, independent of object geometry. **Bottom (Grasp Synthesis):** At inference, 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:

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

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-

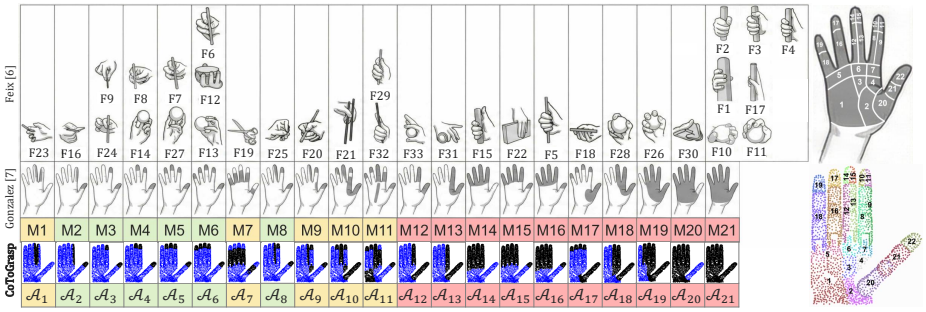

Fig. 3: Semantic Grasp Taxonomy and Contact Mapping. (Left) Correspondences between the classical Feix [11] (F) taxonomy (top), the haptic Gonzalez [12] (M) taxonomy (middle row) and our derived point cloud contact templates $\mathcal{A}_m$ (bottom row). We categorized the 21 templates into three distinct 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 [12], 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 [12] 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. [17, 24] 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 [24], 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 [34] 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, $\Lambda_m \in \{0, \dots, N\}^{N_W}$ 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 [40] 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 [20]. 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 [37] 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 [47] 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 [32]. 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 [17,24,32]. 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 [39]. 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.) ↑		
					t (m)	R (rad)	Q (rad)
DFC [26]	✓	72.15	0.7389	>1800	0.0607	1.424	**0.3579**
GenDexGrasp [24]	✗	71.15	0.5956	14.65	0.0519	1.416	0.2567
DRO-Grasp [43]	✗	63.30	0.6504	1.72	0.0546	**1.515**	0.2892
GOAG [32]	✓	**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 [24, 32, 43] (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 $\tilde{f}_m = \mathbf{TC}_m / \sum_{j=1}^{M} \mathbf{TC}_j$.

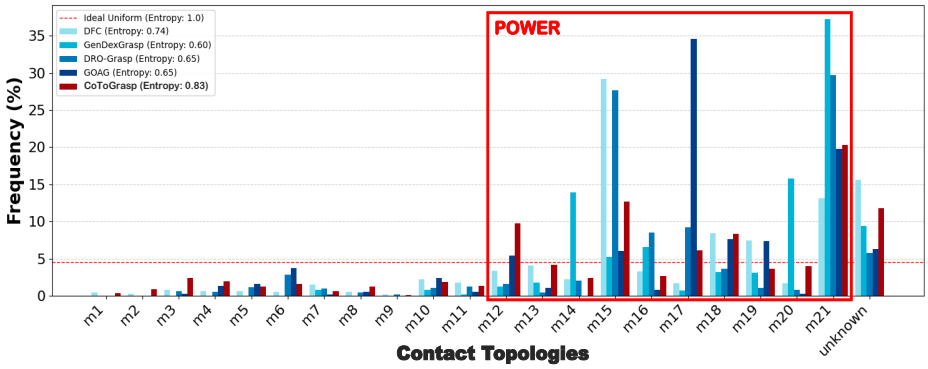

Fig. 4: Functional contact topology distribution across taxonomy-unaware planners. The histogram illustrates the distribution of grasps generated by unconditioned baselines compared to CoToGrasp on the Multidex objects set. The unknown category represents physically stable grasps with unnatural contact 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 [24,26,32,43] 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 [5] on the DexGraspNet [41] test set. After mapping Dexonomy's Feix [11] 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 [5]	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 [5] 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).

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 [5] 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 [5] 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.

Method	SR (%)					
	M2	**M6**	**M11**	**M13**	**M18**	**M21**
Dexonomy [5]	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.

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

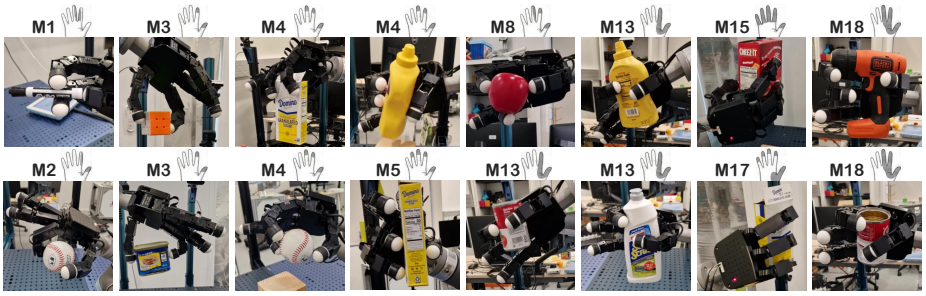

Fig. 5: Real-World kinematic validation. CoToGrasp synthesizes diverse, topology-compliant grasps that are physically executable on a physical Allegro Hand using YCB [4] objects. The target contact topologies (indicated above each frame) demonstrate the physical viability of the generated grasps across 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 [4] 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 n^{o} 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. Brahmbhatt, S., Ham, C., Kemp, C.C., Hays, J.: Contactdb: Analyzing and predicting grasp contact via thermal imaging. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2019)
3. Bullock, I.M., Ma, R.R., Dollar, A.M.: A hand-centric classification of human and robot dexterous manipulation. IEEE transactions on Haptics (2012)
4. 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)
5. Chen, J., Ke, Y., Peng, L., Wang, H.: Dexonomy: Synthesizing all dexterous grasp types in a grasp taxonomy. Robotics: Science and Systems (2025)
6. Chitta, S., Sucan, I., Cousins, S.: Moveit![ros topics]. IEEE robotics & automation magazine (2012)
7. 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)
8. 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)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
10. 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)
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 (2015)

12. Gonzalez, F., Gosselin, F., Bachta, W.: Analysis of hand contact areas and interaction capabilities during manipulation and exploration. IEEE transactions on haptics (2014)
13. 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)
14. Huang, L., Zhang, H., Wu, Z., Christen, S., Song, J.: Fungrasp: Functional grasping for diverse dexterous hands. IEEE Robotics and Automation Letters (2025)
15. 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)
16. 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)
17. 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)
18. 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)
19. 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)
20. 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)
21. 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)
22. Li, H., Mao, W., Deng, W., Meng, C., Fan, H., Wang, T., Osamu, Y., Tan, P., Wang, H., Deng, X.: Multi-grasplllm: A multimodal llm for multi-hand semantic guided grasp generation. arXiv preprint arXiv:2412.08468 (2024)
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 (2024)
24. 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)
25. Liu, M., Pan, Z., Xu, K., Ganguly, K., Manocha, D.: Deep differentiable grasp planner for high-dof grippers. arXiv preprint arXiv:2002.01530 (2020)
26. 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)
27. 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)
28. Lu, Q., Hermans, T.: Modeling grasp type improves learning-based grasp planning. IEEE Robotics and Automation Letters (2019)

29. 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)
30. Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., Hoeller, D., Rudin, N., Allshire, A., Handa, A., et al.: Isaac gym: High performance gpu based physics simulation for robot learning. In: NeurIPS Datasets and Benchmarks (2021)
31. 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)
32. 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)
33. Miller, A.T., Allen, P.K.: Graspit! a versatile simulator for robotic grasping. IEEE Robotics & Automation Magazine (2004)
34. 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)
35. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2017)
36. 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)
37. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. Advances in neural information processing systems (2015)
38. 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)
39. Tversky, A.: Features of similarity. Psychological review, American Psychological Association (1977)
40. 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)
41. 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)
42. 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)
43. 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)
44. Weng, Z., Lu, H., Kragic, D., Lundell, J.: Dexdiffuser: Generating dexterous grasps with diffusion models. IEEE Robotics and Automation Letters (2024)
45. Wu, R., Zhu, T., Lin, X., Sun, Y.: Cross-category functional grasp transfer. IEEE Robotics and Automation Letters (2024)

46. 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)
47. Xu, J., Sun, X., Zhang, Z., Zhao, G., Lin, J.: Understanding and improving layer normalization. Advances in neural information processing systems (2019)
48. Xu, Q., Xu, Z., Philip, J., Bi, S., Shu, Z., Sunkavalli, K., Neumann, U.: Point-nerf: Point-based neural radiance fields. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2022)
49. 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)
50. 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)
51. 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)
52. 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)

Supplementary Material

This supplementary material provides additional technical details to support the main manuscript. Section A details the taxonomy transfer and handprint discretization. Section B outlines the network architecture and hyperparameters. Section C justifies the train-test domain shift quantitatively and qualitatively. Section D provides the mathematical formulations for our validation pipeline and energy-based optimization. Finally, Sections E through K offer extended experimental protocols, metric definitions, and comprehensive qualitative results.

A Applying contact topologies to grippers

Handprint Discretization. To generate the spatial handprints $\mathcal{H}(Q)$ for the training dataset, we compute the forward kinematics for each sampled joint configuration Q. To represent the gripper's active grasping surfaces, we discretize its geometry into a fixed point cloud of $N_H = 2,048$ points. We filter the dense surface points based on their spatial dot products relative to the approach direction, ensuring a uniform, configuration-independent resolution across the functional contact areas. Specifically, let $\mathbf{G} \in \mathbb{R}^{M\times 3}$ represent the matrix of surface normals for a dense set of M candidate points on the gripper's geometry. We define a reference direction vector, $\mathbf{v}_{\text{ref}} \in \mathbb{R}^{3\times 1}$, representing the palm's facing direction (e.g. $\mathbf{v}_{\text{ref}} = [0, -1, 0]^T$ for the Shadow Hand). To isolate the active grasping surfaces, we evaluate the alignment of all candidate points simultaneously by computing the dot products between their normals and the reference direction:

$$\mathbf{s} = \mathbf{G}\mathbf{v}_{\text{ref}} \tag{12}$$

The resulting scalar values in the vector $\mathbf{s} \in \mathbb{R}^{M\times 1}$ are then used to threshold and sample the most relevant contact surfaces, culminating in the final N_H points that form $\mathcal{H}(Q)$. Figure 6 shows the resulting handprints for the Shadow and Allegro hands, alongside the manually defined zone divisions detailed in Section 3.1.

Taxonomy Transfer to Anthropomorphic Grippers. To semantically ground the generated handprints $\mathcal{H}(Q)$, we map established human grasp taxonomies onto the robotic kinematics. We manually partition the active grasping surfaces of both the Shadow Hand and the Allegro Hand into a discrete set of anatomical zones. Specifically, the mechanical links and palm areas are segmented into the $M = 21$ contact topology templates. During the handprint generation process, each sampled point $\mathbf{h} \in \mathcal{H}(Q)$ is assigned to its corresponding zone $\mathcal{A}_k$. Notably, because the Allegro Hand lacks a fifth finger, the contact topology templates for $\mathcal{A}_5$ and $\mathcal{A}_6$ are functionally identical. To prevent redundancy, we explicitly omit $\mathcal{A}_6$ from the Allegro Hand's template set. Ultimately, this taxonomy transfer (illustrated Figure 7) ensures that despite the morphological differences between the grippers, their spatial handprints share a common semantic representation, enabling structured, cross-embodiment comparisons of grasp strategies.

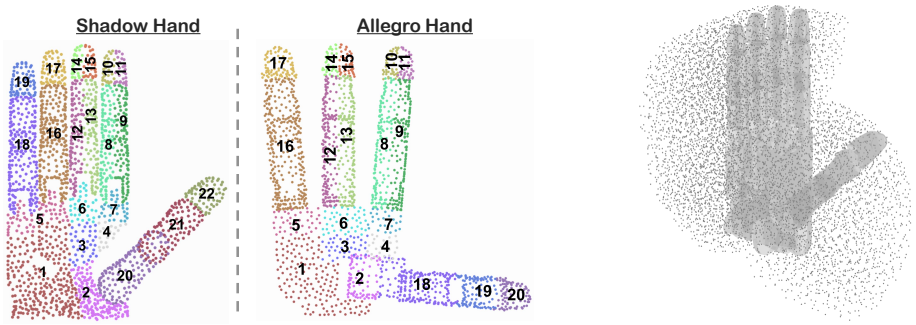

Fig. 6: Handprint areas and workspace constitution. Left: Discretized handprints of the Shadow and Allegro hands, illustrating the manually defined anatomical zone divisions. **Right:** A three-quarter view of the Shadow Hand's workspace.

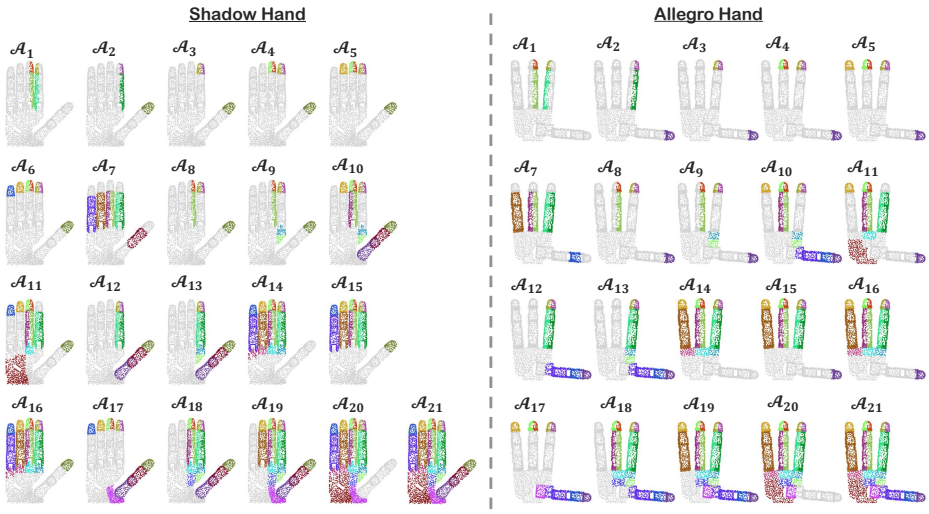

Fig. 7: Taxonomy transfer mapping for anthropomorphic grippers. The handprints of the Shadow Hand (left) and the Allegro Hand (right) are segmented into manually defined, corresponding anatomical zones ($\mathcal{A}_1$–$\mathcal{A}_{21}$).

Workspace Computation. We define the gripper's workspace as an aggregate point cloud representing its reachable surface area. This representation is generated by superimposing 10,000 gripper handprints sampled from our synthetic dataset, effectively capturing the complete interaction space of the fingers and palm independently of any external environment. To construct our fixed-size set of spatial basis points $\mathcal{W} = \{w_j \in \mathbb{R}^3\}_{j=1}^{N_W}$, this dense accumulation is downsampled using Farthest Point Sampling (FPS). This yields a spatially uniform distribution of N_W that comprehensively encapsulate the gripper's kinematic workspace. Figure 6 shows an illustration of the Shadow Hand's workspace.

B Architectural Choices and Implementation Details

To ensure the CoToGrasp framework accurately models the complex physics and semantics of dexterous grasping, several deliberate architectural choices were implemented. This section details the theoretical justifications for these design choices.

Local Geometry Extraction vs. Global Shape Descriptors. Standard 3D generative pipelines often employ global shape descriptors (e.g. standard PointNet [35]) that compress the entire object into a single latent vector. While effective for classification, global compression destroys high-frequency spatial details – such as local curvature and surface normals – which are absolutely critical for identifying physically viable contact patches. For this, we implement a DGCNN [34] encoder using the architecture presented in [43]. Unlike the standard *Edge Convolution* operation that dynamically recomputes K-Nearest Neighbors at every network layer, this architecture utilizes a fixed graph structure throughout, resulting in a "Static Graph CNN". Specifically, we set the neighborhood size to $K = 16$ to capture highly localized geometric information. By utilizing this "Static Graph CNN" architecture, we preserve point-wise local features $\mathcal{F}_{\mathcal{P}}$, ensuring the network reasons over local geometric fidelity while remaining invariant to the global spatial arrangement of the point cloud. Its learned weights successfully generalize across the domain shift – from the structured kinematic handprint seen during training to the highly variable target objects processed during inference.

Features Aggregation during the Workspace Projection. During the feature projection phase onto the canonical workspace $\mathcal{W}$, we utilize a modified k-Nearest Neighbors aggregation. Inspired by scattered data interpolation in neural rendering [48], this local pooling strategy preserves high-frequency details more effectively than dense voxel grids or global pooling. In standard scattered data interpolation, it is common to normalize the aggregated features by the sum of the exponential weights. We intentionally omit this normalization step, instead computing an unnormalized average divided strictly by k (Eq. 5). Because the workspace covers the entire kinematic reach of the gripper, many basis points lie in "empty space," far from the input geometry. By leaving the distance weights unnormalized, they naturally decay to zero for distant neighbors. This conse-

quently suppresses feature activation in empty regions, effectively preventing the network from hallucinating object geometry where none exists.

Persistent Hard Conditioning in Self-Attention. Standard vision transformer architectures [9] (ViT) typically condition the generation process by prepending the condition variable as an isolated global [CLS] token. However, grasp synthesis relies heavily on precise local contact arrangements across thousands of spatial tokens. A single global token is prone to signal dilution across deep attention layers [58]. To circumvent this, we explicitly concatenate the semantic topology embedding $\mathcal{F}_{\mathcal{T}}$ to every single workspace point feature. This point-wise fusion enforces a persistent, "hard" conditioning across the entire spatial domain, ensuring that every local geometric feature is evaluated strictly under the lens of the target functional intent at every layer of the network.
Since the projected workspace features $\mathcal{F}_{\mathcal{W}}$ are processed by the transformer as an unordered sequence of tokens – lacking the explicit 3D coordinates w_i of the original basis points – we inject spatial awareness via sinusoidal positional encodings (PE). This allows the multi-head attention mechanism to recover spatial relationships purely from the fixed basis set indices. Crucially, the network is not provided with explicit contact priors at inference time. Instead, the semantic knowledge of valid contact configurations is acquired implicitly through backpropagation. The supervision signal drives the multi-head self-attention mechanism to discover and amplify geometric correlations between spatially distant points that correspond to stable configurations.

Global Latent Pooling via Set Transformer. Because our architecture explicitly omits a global learnable [CLS] token to preserve dense local conditioning, we cannot simply extract a designated output token (such as the $\mathbf{z}_L^0$ state in standard ViTs) to form our global representation. Consequently, an explicit aggregation strategy is required to synthesize the dense sequence of fused geometric and semantic features into a unified global latent descriptor $\mathcal{Z}$. For this task, we opted against static pooling operators (e.g. max-pooling or average-pooling). Static operators process elements independently, discarding non-extremal data and failing to capture spatial correlations. Instead, we utilize a Set Transformer. The Pooling by MultiHead Attention (PMA) block adaptively weighs input instances based on task relevance, while the subsequent Set Attention Block (SAB) explicitly models pairwise and higher-order interactions. This allows the network to successfully encode the structural co-occurrences and spatial distributions of distant contact patches.

CVAE and Contact-Topology-Conditioned Multi-Modality. A well-documented limitation of standard CVAEs in grasp generation is their tendency to suffer from posterior collapse [15], [54], [55] or blurred predictions when faced with the highly multi-modal nature of general grasping (e.g. the exact same object can be pinched, enveloped, or grasped by the rim). CoToGrasp bypasses this limitation. Because the macroscopic multi-modality is explicitly resolved by the strict topology conditioning (which dictates the exact functional approach), the CVAE is relieved of the burden of modeling drastically different grasp modes. It is only required to model the localized, intra-topology spatial variations of

the contact patches (e.g. slight finger shifts along an edge). For this highly constrained stochasticity, a standard unimodal Gaussian prior $\mathcal{N}(0, I)$ proves to be both mathematically sufficient and computationally highly efficient. The variational lower bound (ELBO) consists of the reconstruction term, a Multi-Class Cross-Entropy loss (L_{recon}) over the workspace points $\mathcal{J} = \{1, \ldots, n_W\}$, and a Kullback-Leibler regularization term:

$$L = L_{\text{recon}} + \beta D_{\text{KL}}(q(z|\mathcal{Z}) \parallel \mathcal{N}(0, I)), \quad L_{\text{recon}} = -\sum_{j \in \mathcal{J}} \mathbf{y}_j \cdot \log(\hat{\mathbf{y}}_j) \tag{13}$$

where β is a weighting parameter balancing latent capacity and reconstruction accuracy, $\mathbf{y}_j$ is the one-hot encoded ground truth derived from $\Lambda_m(j)$ and $\hat{\mathbf{y}}_j$ is the predicted probability vector.

Network Hyperparameters and Training Details. Table 4 summarizes the comprehensive architectural details, layer dimensions, and hyperparameters used to implement and train the CoToGrasp framework. All training procedures were executed across a cluster of 4 Nvidia A100 GPUs, requiring approximately 35 hours.

C Train-Test Domain Shift and Feature Alignment

To demonstrate that our architectural choices enable robust cross-domain generalization – specifically, the transfer from the gripper surface (training domain) to the object surface (inference domain) – we analyze the latent feature alignment across both domains. This evaluation verifies whether our framework successfully bridges the inherent geometric and structural differences between hands and objects.

Quantitative Alignment Analysis. To measure the transfer quality, we extracted intermediate feature representations from both domains and evaluated their alignment at active topological contact points using Average Cosine Similarity. We compared corresponding spatial points (*Matched Pairs*) against random, non-corresponding points (*Random Pairs*) to establish a baseline (Table 5). Initially, the raw point-wise features extracted from the DGCNN ($\mathcal{F}_{\mathcal{P}}$) exhibit a significant domain gap, achieving a poor similarity score of 0.35 for matched pairs and 0.23 for random pairs. This confirms that the raw domain shift remains substantial despite static graph modifications. However, the effectiveness of projecting into our feature-based canonical workspace via kNN aggregation ($\mathcal{F}_{\mathcal{W}}$) is immediately evident: the matched pair cosine similarity jumps to 0.61. This quantitatively proves that the canonical projection successfully disentangles features from the input topology, effectively bridging the domain gap. Finally, introducing the contact topology embedding ($\mathcal{F}_{\mathcal{T}}$) into the Transformer encoder (Φ) further sharpens this representation. The self-attention mechanism increases matched similarity to 0.66 while actively suppressing mismatched, domain-specific geometric noise (reducing random pair similarity from 0.33 to 0.27).

Qualitative Latent Space Visualization. This quantitative improvement is visually corroborated by the t-SNE projection of our canonical workspace features ($\mathcal{F}_{\mathcal{W}}$), illustrated in Figure 8. While global domain distinctions naturally

Module / Parameter	Notation	Value / Size
Training & Optimization		
Batch Size	–	32
Learning Rate	–	10^{-5}
Training Epochs	–	50
KLD Regularization Weight	β	0.1
Attention Factor	α	3.0
Workspace & Geometric Projection		
Workspace Resolution	N_W	$8,192$
Projection kNN	k	5
Aligned Distance Scaling	γ	2.0
Pointwise Feature Extraction		
Local Graph kNN	K	16
Hidden Layer Sizes	–	$[12, 64, 64, 128, 256, 512]$
Output Feature Dimension	$\mathcal{F}_\mathcal{P}, \mathcal{F}_\mathcal{W}$	$1,024$
Conditioning Embeddings		
Topology Embedding Dim.	$\mathcal{F}_\mathcal{T}$	64
Label Embedding (MLP)	$\mathcal{F}_\mathcal{A}$	$[128, 64]$
Self-Attention Modules		
Transformer Encoder Feature Dim.	Φ	1088
Transformer Encoder Blocks	–	4
Transformer Encoder Heads	–	8
Set Transformer (Pooling) Heads	–	8
Global Latent Descriptor Dim.	$\mathcal{Z}$	$1,024$
CVAE & Latent Space		
Latent Encoder Hidden Dim.	–	512
Latent Variable Dimension	ψ	64
AdaLN Decoder Output Classes	$N + 1$	23

Table 4: Implementation Details and Network Hyperparameters for the CoToGrasp framework.

persist – reflecting the inherent macro-structural differences between an anthropomorphic hand and arbitrary objects – the features distinctly interleave within shared manifold clusters at task-relevant regions. This selective alignment visually and mathematically justifies our reliance on the self-attention layers (Φ). The Transformer is necessary to isolate and attend to these domain-invariant contact primitives, resolving non-local spatial dependencies while ignoring irrelevant structural disparities. Ultimately, this ensures that the generative process remains robust regardless of the input surface.

	Raw	DGCNN	Workspace + Attn.
Matched Pairs	**0.35**	**0.61**	**0.66**
Random Pairs	0.23	0.33	0.27

Table 5: Feature Alignment (Cosine Similarity).

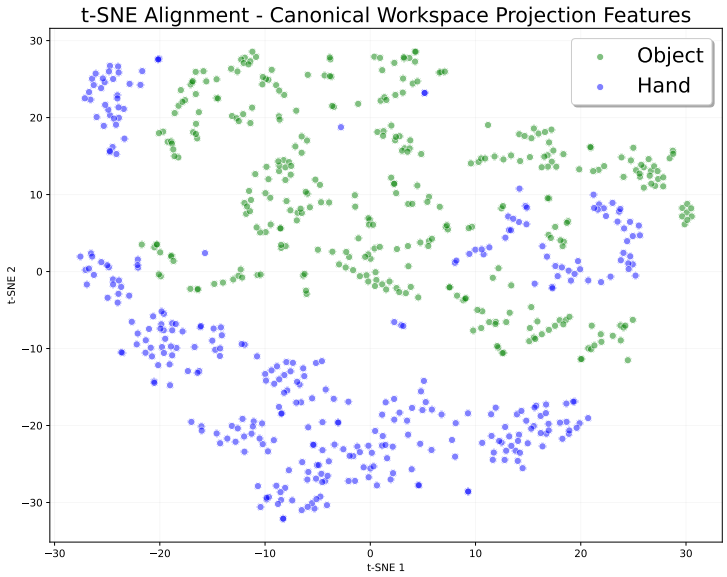

Fig. 8: t-SNE visualization of features after canonical workspace projection.

D Validation Pipeline and Energy Optimization Details

This section details the post-inference validation and optimization pipeline, which is designed to filter out physically and topologically invalid predictions before finalizing the hand's kinematics. To ensure both semantic correctness and physical feasibility, our framework processes the generated contact templates through a sequential validation strategy: a label-consistency check, a preliminary force-closure evaluation, and an energy-based joint optimization. The average runtime per grasp evaluates to 110 ms, with the computational cost broken down as follows: pose initialization (2.8 ms), forward network inference (0.3 ms), label-consistency verification (42.4 ms), force-closure computation (14.7 ms), and the final joint configuration optimization (50.7 ms).

Label-Consistency Check Formulation. Let $\mathcal{U}(\Lambda) = \{\Lambda(j) \mid \Lambda(j) \neq 0\}$ be an operator that extracts the set of unique, active semantic zone identifiers from a given contact template. To prevent the generation of ill-posed grasps, we define the validation function $V(\Lambda_m, \hat{\Lambda}_m)$ as a strict binary check on the symmetric difference between the extracted zone sets of the ground truth and predicted templates::

$$V(\Lambda_m, \hat{\Lambda}_m) = \mathbb{I}(|\mathcal{U}(\Lambda_m) \Delta \mathcal{U}(\hat{\Lambda}_m)| \leq 1) \tag{14}$$

where $\mathbb{I}$ is the indicator function and Δ is the symmetric difference operator defined as $\mathcal{U}(\Lambda_m)\Delta\mathcal{U}(\hat{\Lambda}_m) = (\mathcal{U}(\Lambda_m) \setminus \mathcal{U}(\hat{\Lambda}_m)) \cup (\mathcal{U}(\hat{\Lambda}_m) \setminus \mathcal{U}(\Lambda_m))$. Setting the tolerance threshold to 1 allows for a maximum deviation of one missing or hal-

lucinated contact zone (e.g. a missing palm contact in a power grasp) while successfully rejecting fundamentally distinct topological failures.

Force-Closure Formulation. Unlike methods trained on pre-validated grasp datasets, our generative approach requires an explicit assessment of grasp stability. To ensure the inferred contact points can theoretically yield a stable grasp before performing the computationally expensive joint configuration optimization, we evaluate the Force-Closure condition using the explicit spatial coordinates of the active workspace points.

- **Contact Modeling:** We first extract the subset of workspace points assigned a valid contact label: $\hat{\mathcal{C}}(\mathcal{W}) = \{w_j \in \mathcal{W} \mid \hat{\Lambda}_m(j) \neq 0\}$. We then group these active points by their predicted semantic zone indices. To enforce a unique contact location per required zone, we compute the spatial barycenter of each point cluster and project it onto the object's surface $\tilde{\mathcal{O}}$, yielding a set of discrete contact locations b_i. We assume a Coulomb friction model with a friction coefficient of $\mu = 0.3$. Because this estimation is performed entirely within the canonical gripper's reference frame to assess geometric feasibility, we do not factor in gravity or the object's mass at this stage.
- **Force-Closure Computation:** We compute the grasp wrench space, defined as the convex hull of all possible wrenches generated by unit contact forces at locations b_i. The total wrench is given by:

$$d = \sum_{i=1}^{k} d_i = \sum_{i=1}^{k} G_i f_i$$

  where G_i is the partial grasp matrix for contact b_i, and f_i represents the primitive force vectors along the edges of the friction cone, normalized such that $\|f_i\| = 1$. We verify the force-closure condition by checking if the origin of the wrench space lies strictly within the interior of this convex hull.
- **Theoretical vs. Physical Quality:** It is important to note that this analytical step strictly validates the theoretical stability potential of the semantic contact configuration using a simplified barycenter model. The final, true physical grasp quality is determined during the energy-based optimization, which forces the gripper's complex kinematics to align with the full, dense 3D points of $\hat{\mathcal{C}}(\mathcal{W})$ rather than just these idealized discrete barycenters.

Joint Configuration Optimization Formulation. During the final optimization phase, the joint configuration Q^* is retrieved by minimizing a composite energy function. The weighting factors (λ) balancing the individual energy terms are empirically set to ensure stable convergence. The individual energy terms and their respective weights are defined as follows:

- **Contact Distance (E_{dist}, $\lambda_d = 1.0$):** To encourage precise alignment between the predicted spatial contact points $\hat{\mathcal{C}}(W)$ and the gripper's geometry, we minimize the squared Euclidean distance between each target contact point and the closest point on its assigned kinematic link:

$$E_{dist} = \sum_{p \in \hat{\mathcal{C}}(W)} \min_{h \in \mathcal{H}_{l(p)}(Q)} \|p - h\|_2^2 \tag{15}$$

where $\mathcal{H}_{l(p)}(Q)$ represents the subset of handprint points belonging to the specific link $l(p)$ assigned to the contact point p.

– **Object Penetration (E_{pen}, $\lambda_p = 0.5$):** To ensure physically plausible grasps, we explicitly penalize gripper points that penetrate the target object's volume. Utilizing the object's Signed Distance Field (SDF), denoted as $\Phi_{\mathcal{O}}(\cdot)$, this term is formulated as:

$$E_{pen} = \sum_{h \in \mathcal{H}(Q)} \max(0, -\Phi_{\mathcal{O}}(h)) \tag{16}$$

– **Self-Penetration (E_{spen}, $\lambda_s = 0.01$):** We prevent self-collisions between different fingers or parts of the robotic hand by enforcing a minimum spatial safety threshold τ (set to 0.025 m) between disjoint kinematic links i and j:

$$E_{spen} = \sum_{i,j} \max(0, \tau - \text{dist}(k_i(Q), k_j(Q)))^2 \tag{17}$$

where $k_i(Q)$ denotes the spatial centroid of the bounding box for link i in a given configuration Q.

– **Joint Limits (E_j, $\lambda_j = 1.0$).** We strictly constrain the optimized joint angles to remain within the physical hardware's kinematic limits, defined by $[Q_{min}, Q_{max}]$:

$$E_j = \| \max(0, Q - Q_{max}) \|_2^2 + \| \max(0, Q_{min} - Q) \|_2^2 \tag{18}$$

– **Repulsive Energy Term (E_{rep}, $\lambda_r = 0.01$):** During the energy-based joint optimization, we aim to align the gripper to the predicted contacts while strictly preventing unintended parts of the hand from resting on the object, which would violate the requested contact topology. To achieve this, we introduce the repulsive energy term E_{rep}, which penalizes object proximity for all gripper handprint points belonging to non-active regions:

$$E_{rep} = \sum_{h \in \mathcal{H}(Q) \setminus \mathcal{H}_{l(p)}(Q)} \max(0, \delta - \text{dist}(h, \tilde{\mathcal{O}})) \tag{19}$$

We enforce a minimum safety margin of $\delta = 0.005$ m.

E Topology-Conditioned Pose Sampling Heuristic

As discussed in the main text, CoToGrasp treats the global 6D grasp pose as an external prior. To autonomously evaluate the framework without relying on external task planners, we sample diverse candidate object poses $(R, \mathbf{t})^{-1}$ using a topology-conditioned heuristic tailored to the specific active zones of the requested grasp template. Figure 9 illustrates the following sampling strategy for three distinct contact topologies.

Palm-Constrained Topologies. For power grasps and contact topologies requiring palm contact, we adopt the spatial heuristic from [24, 26, 32]. We uniformly sample approach translations and orientations on the object's convex

hull, directed toward its volumetric center. The object is then transformed into the canonical gripper frame via the inverse transformation $(R, \mathbf{t})^{-1}$.

Distal and Precision Topologies. For grasps explicitly excluding the palm (e.g. fingertips only), standard convex hull sampling often yields kinematically unreachable configurations. Instead, we sample the object pose $(R, \mathbf{t})^{-1}$ directly within a restricted kinematic sub-workspace of the target template's active zones. To ensure the object is placed in a feasible region away from the palm, we systematically truncate the sub-workspace using three geometric constraints: (1) a radial distance threshold from a predefined workspace center to exclude out-of-reach boundary points, (2) a directional half-space filter ensuring points lie strictly in front of the palm's normal axis, and (3) a minimum height threshold along the z-axis to avoid the lower hand geometry. Furthermore, to guarantee topological validity and prevent trivial local minima during the energy-based joint optimization, we explicitly enforce a strict 1 cm clearance margin between the object geometry and the palm.

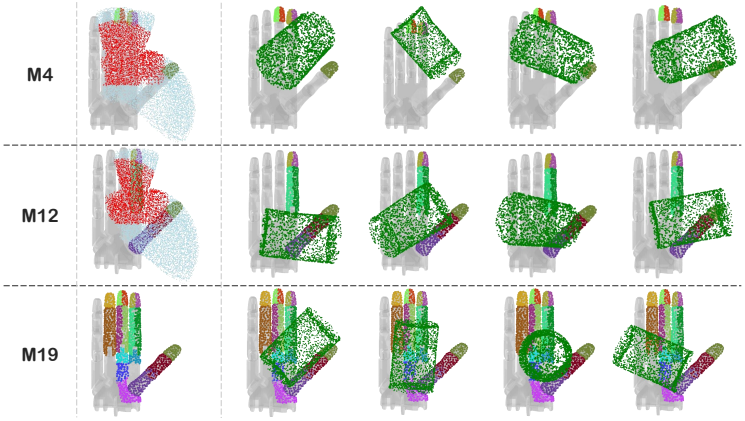

Fig. 9: Topology-Conditioned Pose Sampling. For specific grasp topologies (such as M4 and M12), the initial object pose is sampled within a restricted kinematic region. **Middle:** The template's full active sub-workspace is shown in light blue, while the truncated sub-workspace – filtered for reachability and palm clearance – is highlighted in red. **Right:** Examples of the initialized object point cloud $\tilde{\mathcal{O}}$ (green) successfully placed within this feasible region after the sampled spatial transformation.

F Automated Grasp Classification Metric

This section details the mathematical formulation of our contact topology recognition metric.

Contact Extraction and Zone Mapping. First, we extract the subset of gripper points in contact with the object, $\mathcal{C}_{\text{grip}}(\mathcal{H}(Q))$, using the aligned distance formulation with an empirically chosen proximity threshold of $\epsilon = 0.8$. To compare

these physical contacts against the taxonomy $\mathcal{T}$, we map the raw contact points back to their semantic zone identifiers. Let $I_{\text{obs}} = \{i \in \mathcal{I} \mid \mathbf{h}_i(Q) \in \mathcal{C}_{\text{grip}}(\mathcal{H}(Q))\}$ be the indices of the gripper points currently touching the object. Using the surjective mapping ζ, we define the observed active zones (A_{obs}) as the set of unique physical regions engaged in the grasp, and the ground-truth required zones (A_m) extracted from the target template $\mathcal{A}_m$ (explicitly excluding the label 0):

$$A_{\text{obs}} = \{\zeta(i) \mid i \in I_{\text{obs}}\} \qquad \text{and} \qquad A_m = \{\mathcal{A}_m(i) \mid i \in \mathcal{I}\} \setminus \{0\} \tag{20}$$

Asymmetric Tversky Index. We evaluate the match between the generated grasp and a specific contact topology m by computing a similarity score s_m based on the Tversky index [39]:

$$s_m = \frac{|A_{\text{obs}} \cap A_m|}{|A_{\text{obs}} \cap A_m| + w_1|A_{\text{obs}} \setminus A_m| + w_2|A_m \setminus A_{\text{obs}}|} \tag{21}$$

where $|\cdot|$ denotes set cardinality. We set the weighting parameters to $w_1 = 2.0$ and $w_2 = 0.5$. This asymmetric penalization, determined empirically, is critical: it punishes extra contact regions ($|A_{\text{obs}} \setminus A_m|$), which typically denote clumsy or unintended enveloping power grasps, much more strictly than missing contacts ($|A_m \setminus A_{\text{obs}}|$).
Any grasp failing to reach a similarity threshold of $s_m \geq 0.5$ is strictly classified as 'unknown'. Given our asymmetric Tversky weights, w_1 and w_2, 0.5 represents the mathematical tipping point where the number of correctly aligned contact zones strictly outweighs the heavily penalized hallucinated contacts (e.g., a failed precision pinch collapsing into a power grasp), ensuring robust rejection of degenerated topologies.

G Standard Evaluation Metrics Protocol

As mentioned in the main text, CoToGrasp utilizes the standard evaluation protocols widely established in [24, 26, 32, 43] to measure baseline physical performance.
Success Rate (SR). We evaluate the physical stability of the synthesized grasps using the Isaac Gym simulator [30]. The target object (standardized to a mass of 100 g) and the gripper are initialized in the optimized configuration Q^*. We sequentially apply external perturbations on the object along the $\pm x, y, z$ axes for one second each. A grasp is considered successful if the object's translation deviates by less than 2 cm from its initial pose. Furthermore, any grasp exhibiting severe interpenetration (contact forces exceeding 500 N) is strictly classified as a failure to penalize unrealistic physical states.
It is crucial to note that the theoretical force-closure validation performed during inference serves only as a rapid geometric estimation based on idealized point contacts. It offers no absolute guarantees regarding final physical stability, as the Isaac Gym simulation rigorously tests the complex reconciliation of these points against the complete object geometry, strict kinematics, and the torque

saturation limits of the gripper's simulated actuators when compensating for external perturbation forces.

Generation Speed. We report the average computational time (in seconds) required to generate a single grasp, calculated over 100 generation attempts. This metric strictly isolates the network inference and the energy-based joint optimization time, explicitly excluding the heavy physics simulation overhead of Isaac Gym.

Spatial Diversity. We quantify generative spatial diversity by calculating the standard deviation of the gripper translation ($\mathbf{t}$), orientation (R), and joint values (Q). Because the initial global poses $(R, \mathbf{t})$ are derived from our topology-conditioned sampling heuristic, the diversity in $\mathbf{t}$ and R directly reflects the method's spatial reachability and adaptability across various object geometries.

H Detailed Experimental Protocols and Baselines

To ensure a rigorous and fair evaluation, this section provides the extended protocols and baseline configurations.

Taxonomy-Unaware Evaluation Protocol. To evaluate the inherent functional bias of unconditioned grasp planners, we compare CoToGrasp against four state-of-the-art baselines: DFC [26], GenDexGrasp [24], DRO-Grasp [43], and GOAG [32]. For this experiment, we utilize the test subset from the MultiDex dataset, comprising 10 objects from ContactDB [2] and YCB [4]. Each baseline was tasked with generating exactly 100 grasps per object. To ensure a fair comparison, the baselines are deployed as follows:

- **DFC:** as DFC does not require training, we utilize pre-generated grasps from the CMapDataset. These poses were originally synthesized via DFC and subsequently refined by [24] through a post-filtering process to ensure geometric validity.
- **GenDexGrasp:** We use the official pre-trained checkpoint (trained on multi-hand data) with default hyperparameters. We first infer the contact maps for the test objects, which are then passed into the authors' optimization framework to derive the final grasp poses.
- **DRO-Grasp:** We select the most performant variant, which includes configuration-invariant pre-training. We retain all default hyperparameters but explicitly deactivate their simple grasp controller to ensure a fair, purely kinematic comparison.
- **GOAG:** We train GOAG from scratch on the Shadow Hand kinematics using the default hyperparameters, and generate grasps following the standard inference pipeline.

All successfully generated, physically stable grasps were subsequently fed into our automated classification pipeline (Sec. 4.1, Supp. Mat. F) to evaluate their Topology Compliance (**TC**), Entropy and Spatial Diversity.

Because these baselines do not take a functional condition as input, this test strictly isolates their natural generative distribution, exposing the severe mode

collapse toward power grasps driven by standard physics-based optimization. Furthermore, while CoToGrasp achieves a higher **TC** than the baselines, the absolute scores remain relatively low across all methods. This shared ceiling suggests that **TC** is heavily bottlenecked by the specific object set utilized; the limited geometric diversity of these test objects naturally restricts the subset of physically viable grasps, making certain functional topologies kinematically unattainable regardless of the generative method.

Taxonomy-Aware Evaluation Protocol. To evaluate the precise functional control of CoToGrasp, we conducted an extensive comparison against Dexonomy [5] on the full test set of the DexGraspNet [41] dataset, which comprises 1,126 geometrically diverse objects. We selected Dexonomy [5] as our sole baseline for this experiment, as it currently stands as the first and only state-of-the-art framework capable of taxonomy-conditioned generative grasp synthesis. While other recent methods tackle dexterous grasping, they rely on assumptions orthogonal to our object-agnostic setting. For instance, functional grasp transfer and retargeting methods, such as FunGrasp [14] and others [53, 57], operate in a fundamentally different problem space; they inherently require explicit human grasp demonstrations or a dense spatial prior—such as a source human hand mesh—to initialize their optimization. Furthermore, unconditioned frameworks like AnyDexGrasp [10] lack semantic topology control, while methods like OmniDexVLG [51] rely on 2D VLM supervision and are currently publicly unavailable. In contrast, CoToGrasp and Dexonomy [5] function as true grasp planners, synthesizing functionally compliant grasps from scratch relying purely on raw object geometry and a discrete semantic label.

Taxonomy Mapping and Grouping: Dexonomy [5] natively outputs grasps categorized under the classic Feix taxonomy [11]. To ensure a direct one-to-one semantic comparison, we mapped their analytically computed dataset to our contact-based topologies (illustrated in Figure 3 of the main paper). Crucially, this evaluation space ensures an unbiased comparison. By focusing exclusively on contact regions, the Feix taxonomy naturally reduces to our adopted Gonzalez framework (e.g., F12/F13 $\rightarrow$ M6). While Dexonomy couples joint configurations with contact semantics, CoToGrasp relies strictly on contact topologies. Because our metrics (**TC**, $\mathbf{H}_{TC}$) measure semantic compliance purely through these spatial contact assignments, the two taxonomies are structurally equivalent for this assessment. Finally, we explicitly grouped the evaluated grasps into three functional categories to analyze performance across different dexterity levels: *Power* grasps (enveloping, high stability), *Precision* grasps (fingertip manipulation, low stability), and *Object-Specific* grasps (tool use).

Handling Geometric Incompatibility: A critical challenge in large-scale grasp evaluation is that not all contact topologies are geometrically possible on all objects (e.g. it is physically impossible to execute a tiny fingertip precision pinch on a massive, smooth sphere). If an object cannot support a specific contact topology, penalizing the network for failing to generate it skews the metric and misrepresents the model's actual generative capability.

To ensure a strictly fair comparison, we implemented a geometric compatibil-

ity filter. Any object-topology pair that yielded a 0% success rate across all generation attempts was deemed fundamentally geometrically incompatible and explicitly excluded from the final average calculations.
Following this rigorous filtering process, CoToGrasp successfully covers an average of 80.14% of the objects per requested contact topology, which closely and fairly matches Dexonomy's coverage of 81.05%. Consequently, the Success Rate and Topology Compliance metrics reported in Table 2 strictly reflect CoToGrasp's superior generative control on valid geometries, rather than an artifact of object selection.

I Topology Compliance Analysis and Simulation Bottlenecks

To further understand the discrepancy between requested and effective contact topologies observed in the main paper, we analyze CoToGrasp's **TC** at different stages of the generation pipeline (Table 6).

Label Consistency	**Eval. Isaac**	$\mathbf{H}_{SR}$	**TC (%)**	$\mathbf{H}_{TC}$
✓	✓	0.96	17.18	0.84
✓	✗		21.92	0.53
✗	✓	0.94	14.45	0.81
✗	✗		19.34	0.50

Table 6: Topology compliance ablation.

Value of Topological Filtering. Comparing the top and bottom halves of the table demonstrates the value of topological filtering. By rejecting geometrically incompatible templates before the energy-based optimization, the pipeline prevents the optimizer from coercing the hand into unnatural configurations that would ultimately fail in simulation, thereby improving the overall semantic quality of the successful grasps.
The Physics-Based Metric Drop. The most striking shift occurs between the pre-physics (✗ Eval. Isaac) and post-physics evaluations (✓ Eval. Isaac). Prior to the Isaac Gym simulation, the pipeline achieves its highest absolute **TC** but exhibits a remarkably low semantic entropy ($\mathbf{H}_{TC}$). This indicates that the purely geometric, energy-based optimizer frequently converges on configurations that technically satisfy the target contact masks but lack true physical stability. Once subjected to gravity and external perturbations, a significant portion of these superficial grasps naturally fails. While this physics-based filtering prunes weak configurations and drastically rebalances the distribution (restoring $\mathbf{H}_{TC}$), it also causes a drop in the raw **TC**. This drop exposes a sensitivity to the rigid definition of our similarity score formula. During simulation, fingers naturally

shift slightly along the object's local curvature to achieve a stable force-closure. Because our metric enforces strict mathematical boundaries, these minor, functionally viable adjustments often lead to misclassifications. Specifically, despite the presence of a repulsive term in the optimization energy function, certain phalanges cannot be pushed away from the object surface due to inherent kinematic constraints, such as fixed finger lengths or joint limits. In these scenarios, the optimization weighting factors privilege the primary contact required by the topology over the repulsive term, leading to incidental contacts. This phenomenon frequently results in intended precision pinches (M2 or M3) being penalized and reclassified as M12 due to an adjacent phalanx resting too closely to the surface – as illustrated in Figure 10. Consequently, while the final generated grasps are physically stable, this rigid metric provides an incomplete picture of functional success, as it fails to capture whether the intended core contacts were actually achieved.

Fig. 10: Illustration of metric-induced misclassification. While CoToGrasp strictly respects the target contact topology, the kinematic optimization may result in **incidental contacts** where an adjacent phalanx rests against the object surface. Despite the repulsive term in our optimization energy function, these phalanges often cannot be pushed away due to inherent kinematic constraints. For instance, the left grasp illustrates an intended M3 pinch reclassified as M12, while the right shows an M4 grasp reclassified as M15. Although the **intended contacts** are successfully achieved and the grasps remain physically stable, these incidental contacts trigger a strict reclassification by our automated pipeline.

Global Topological Distribution. Beyond the specific pipeline bottlenecks, Figure 11 provides a comprehensive visual breakdown of the attempted versus effective topologies across all 21 topologies. As illustrated, while both methods experience the aforementioned metric-induced shifts during simulation, CoToGrasp maintains a substantially more balanced and faithful distribution of functional grasps (**TC** = 17.18%) compared to the Dexonomy baseline (**TC** = 14.28%).

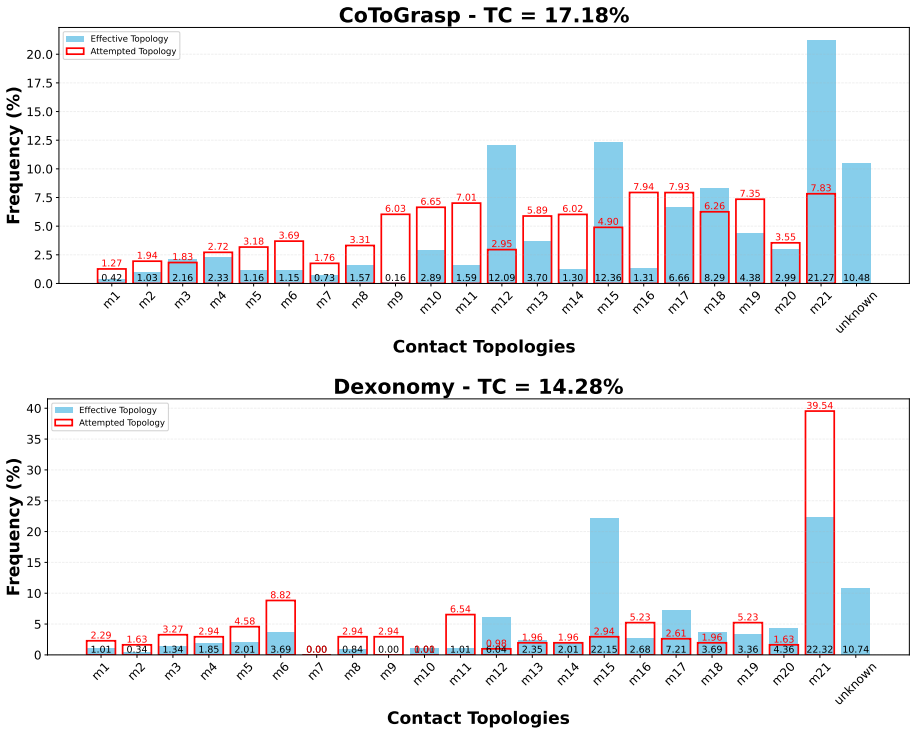

Fig. 11: Histogram comparing the frequencies of **effective topologies** and **attempted topologies** (as defined Sec. 4.2) among all stable grasps generated by CoToGrasp (top) and Dexonomy [5] (bottom).

Object-Level Geometric Complexity. To quantify geometric difficulty and evaluate the inherent trade-off between strict semantic control and geometric flexibility, we analyzed performance grouped by object convexity ($c = V_{obj}/V_{hull}$). As detailed in Table 7, we define *SR Retained* as the comparison of Success Rates (SR) between non-convex ($c < 0.4$) and convex ($c > 0.9$) objects. When transitioning to these challenging geometries, CoToGrasp demonstrates remarkable robustness, retaining 58.72% of its physical SR. In contrast, Dexonomy's rigid templates retain only 37.42% of their performance, and the unconditioned GOAG drops to 28.09%. This performance gap widens on objects with severe concavities ($c < 0.4$, representing roughly 4% of the dataset). On these hardest geometries, CoToGrasp achieves an 18.30% SR, outperforming Dexonomy's 12.05%. Strikingly, CoToGrasp's semantic compliance (**TC**) actually increases to 19.17% on these objects, whereas Dexonomy's **TC** collapses to 8.94%. These results prove that our contact-centric approach thrives on leveraging complex local features to anchor semantics. While strict semantic control typically limits geometric flexibility, CoToGrasp pushes this boundary significantly further than planners relying on rigid, joint-coupled templates, which systematically fail or abandon the semantic query entirely on complex shapes.

Object Complexity	**Metric**	**CoToGrasp**	**Dexonomy [5]**
Convex → Non-Convex	SR Retained	58.72%	37.42%
Severe Concavities	SR	18.30%	12.05%
($c < 0.4$, $\sim 4\%$ data)	TC	19.17%	8.94%

Table 7: Object-Level Analysis. Evaluating performance across geometric complexity ($c = V_{obj}/V_{hull}$). CoToGrasp shows superior retention of both physical stability (SR) and semantic compliance (TC) on challenging non-convex objects.

J Real-World Experiments

Execution Protocol: For a given synthesized grasp $(R, \mathbf{t}, Q)$, we implemented a strict execution pipeline. To ensure collision-free trajectories and safe hardware operation, all motions were initially planned and validated within a ROS2 digital twin using the MoveIt [6] motion planning framework. The execution sequence begins by computing an approach pose, defined by translating the target 6D pose $(R, \mathbf{t})$ backward by 10 cm along the normal vector of the gripper's palm. The robot first navigates to this approach pose with the fingers fully extended. Subsequently, the arm linearly interpolates to the target spatial pose $(R, \mathbf{t})$, and the fingers are actuated to converge on a squeezed configuration Q_s^* based on the optimized joint configuration Q^*. Q_s^* is computed such as finger links involved in the grasps are closer to the object's center of mass, similarly as [5, 43]. To empirically verify the physical stability of the grasp against gravity, the

manipulator lifts the object 15 cm directly above the tabletop, holds it statically for 3 seconds, and finally opens the hand to release the object.

Discussion and Limitations. The execution of synthesized precision grasps on physical hardware exposes the fundamental mismatch between deterministic kinematic planning and the stochastic physics of mechanical contact. While standard position-control frameworks (e.g. OMPL pipelines in MoveIt) are effective for coarse manipulation, they fail to maintain the integrity of complex contact topologies. We observed that relying on a simple linearly interpolated squeezed configuration Q_s^* is critically insufficient; because Q_s^* is derived from geometric distances to the object's barycenter, digits move at uniform velocities regardless of external reaction forces. This inevitably results in asynchronous contact, where leading fingers strike the surface milliseconds early, generating unbalanced moments that displace the object before force-closure is achieved. Consequently, physical contact points drift from predicted optimal zones, misaligning force vectors with the intended functional semantics. Furthermore, current evaluation benchmarks – often restricted to tabletop environments – contradict the objective of omnidirectional grasp synthesis. Such environments artificially inflate the success of grasps that may only be viable in a bimanual setup or when the object is presented in a specific, pre-constrained 6D pose. To maintain fidelity to simulated topologies, future work must transition toward contact-aware interaction paradigms, such as hybrid force/position control or tactile-reactive policies [56].

K Qualitative Results: CoToGrasp Visualization

Figure 12 and Figure 13 present qualitative results across a diverse set of YCB and DexGraspNet objects, highlighting the framework's ability to strictly adhere to the intended contact topology. Rather than merely reaching for the object's center of mass, the synthesized configurations demonstrate precise alignment between the fingers and specific semantic contact zones. As shown in Figure 13, when grasps are grouped by their topological identifiers (M1-M21), the planner consistently recovers the prescribed contact manifolds. This adherence ensures that the resulting configurations maintain the intended grasp type – whether a delicate fingertip pinch or a complex multi-digit wrap – effectively preserving the functional semantics of the interaction across both convex and non-convex surfaces.

Supplementary References

53. Antotsiou, D., Garcia-Hernando, G., Kim, T.K.: Task-oriented hand motion retargeting for dexterous manipulation imitation. In: Proceedings of the European conference on computer vision (ECCV) workshops (2018)
54. Fu, H., Li, C., Liu, X., Gao, J., Celikyilmaz, A., Carin, L.: Cyclical annealing schedule: A simple approach to mitigating kl vanishing. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2019)

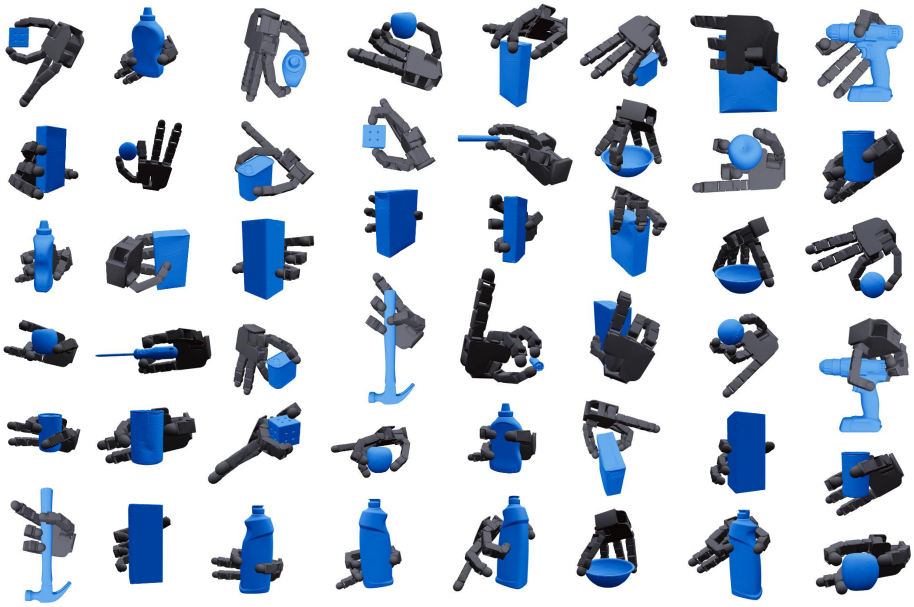

Fig. 12: Qualitative Synthesis Gallery on the Allegro Hand. Synthesized grasp configurations for a diverse subset of the YCB object dataset.

55. He, J., Spokoyny, D., Neubig, G., Berg-Kirkpatrick, T.: Lagging inference networks and posterior collapse in variational autoencoders. In: International Conference on Learning Representations (2019)
56. Jahanshahi, H., Zhu, Z.H.: Review of machine learning in robotic grasping control in space application. Acta Astronautica (2024)
57. Yang, L., Li, K., Zhan, X., Wu, F., Xu, A., Liu, L., Lu, C.: Oakink: A large-scale knowledge repository for understanding hand-object interaction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2022)
58. Zhou, D., Kang, B., Jin, X., Yang, L., Lian, X., Jiang, Z., Hou, Q., Feng, J.: Deepvit: Towards deeper vision transformer. arXiv preprint arXiv:2103.11886 (2021)

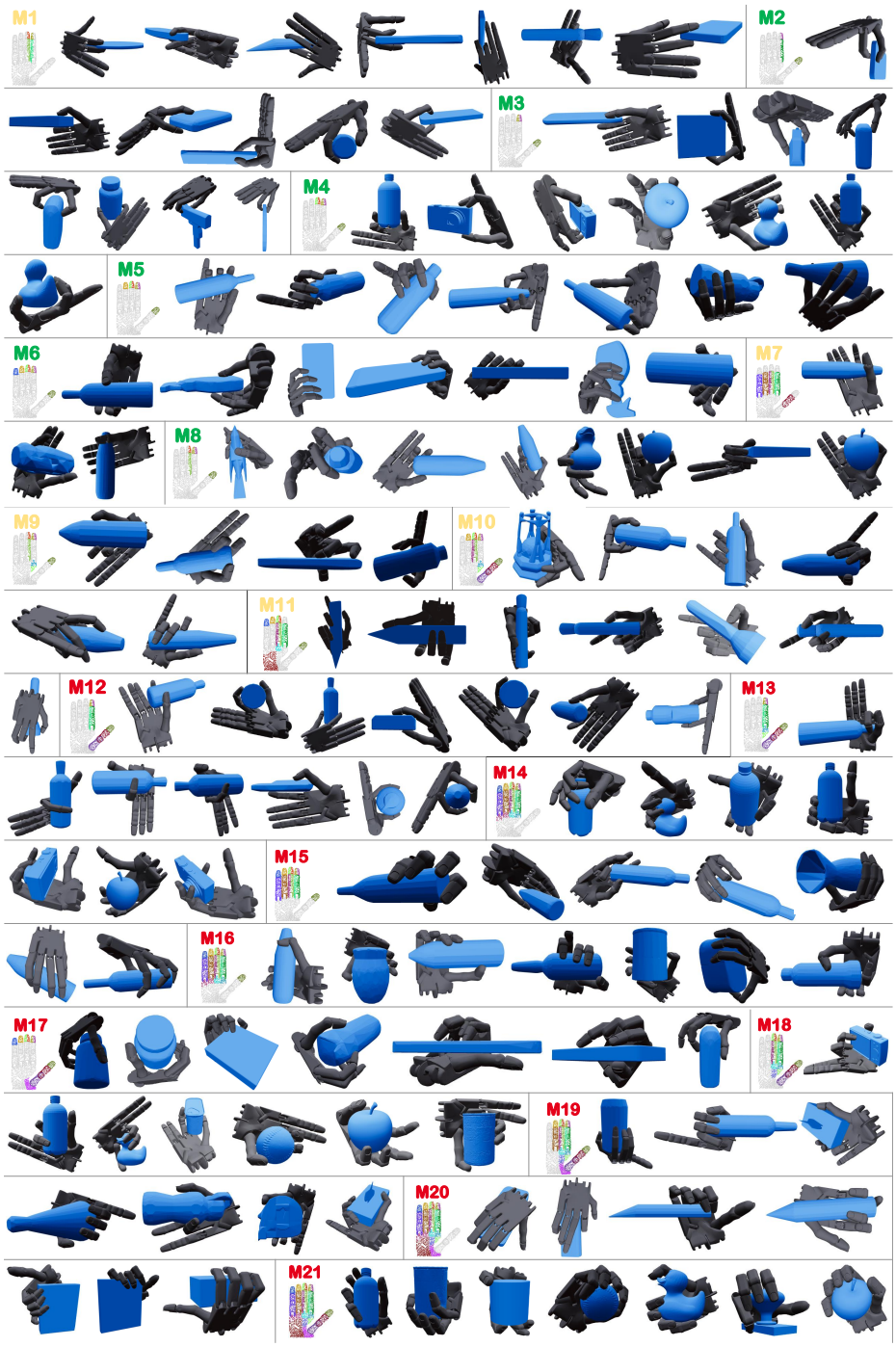

Fig. 13: Topological Clustering of Shadow Hand Grasps. Grasps grouped by contact topology (M1-M21), demonstrating consistent semantic alignment across varied object classes.