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.

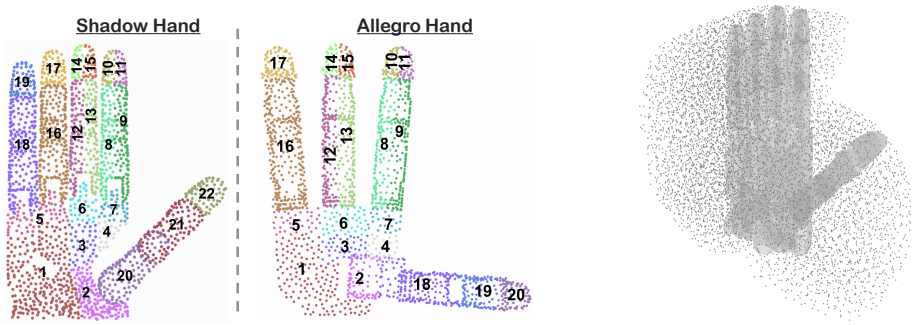

Fig. 6: Handprint areas and workspace constitution. Left: Discretized handprints of the Shadow and Allegro hands, illustrating the manually defined anatomical zone divisions. **Right:** A three-quarter view of the Shadow Hand's workspace.

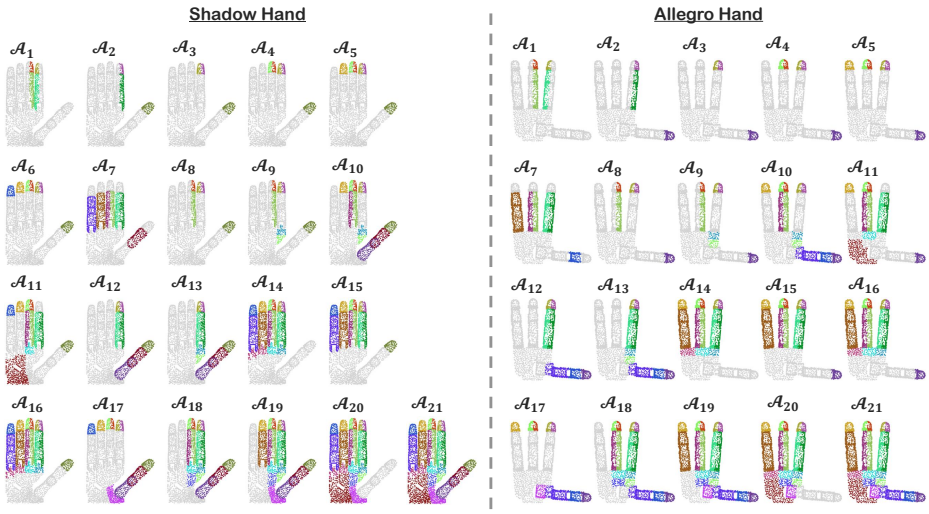

Fig. 7: Taxonomy transfer mapping for anthropomorphic grippers. The handprints of the Shadow Hand (left) and the Allegro Hand (right) are segmented into manually defined, corresponding anatomical zones ($\mathcal{A}_1$–$\mathcal{A}_{21}$).

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).

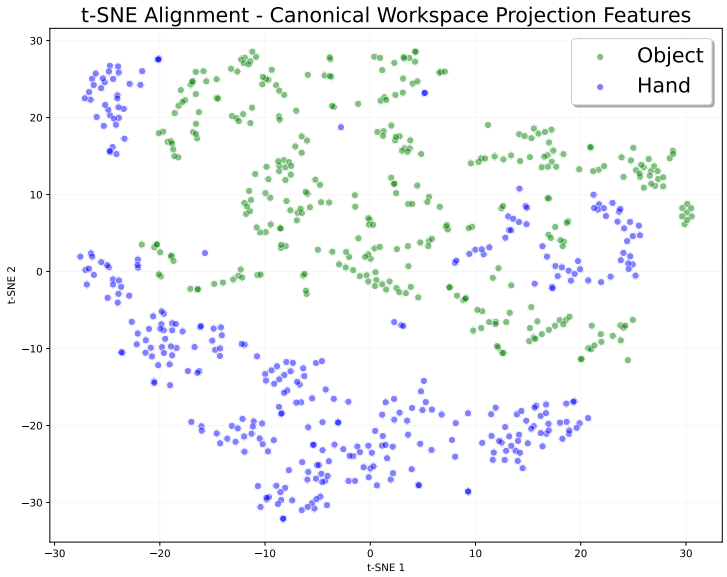

Fig. 8: t-SNE visualization of features after canonical workspace projection.

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.

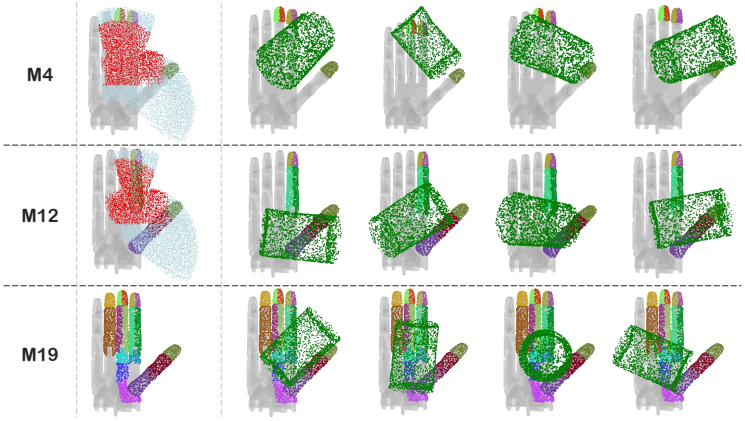

Fig. 9: Topology-Conditioned Pose Sampling. For specific grasp topologies (such as M4 and M12), the initial object pose is sampled within a restricted kinematic region. **Middle:** The template's full active sub-workspace is shown in light blue, while the truncated sub-workspace – filtered for reachability 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%).

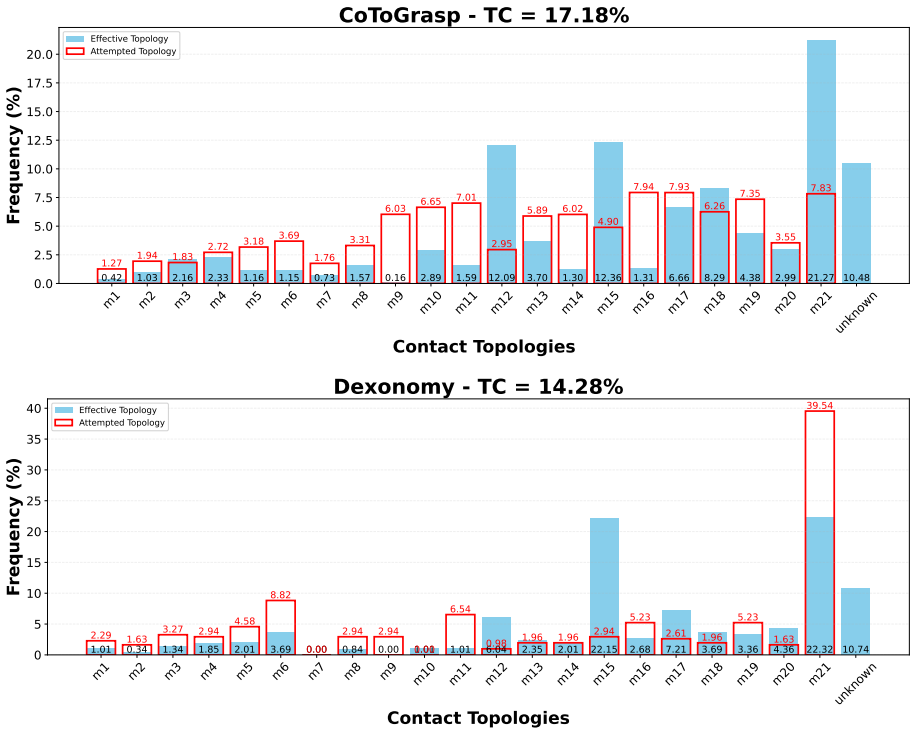

Fig. 11: Histogram comparing the frequencies of **effective topologies** and **attempted topologies** (as defined Sec. 4.2) among all stable grasps generated by CoToGrasp (top) and Dexonomy [5] (bottom).

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)

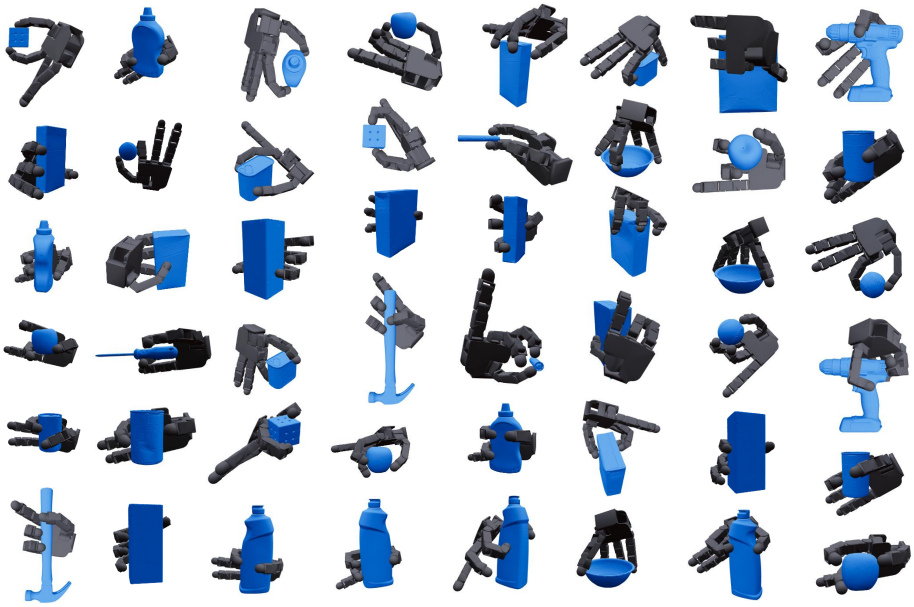

Fig. 12: Qualitative Synthesis Gallery on the Allegro Hand. Synthesized grasp configurations for a diverse subset of the YCB object dataset.

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)

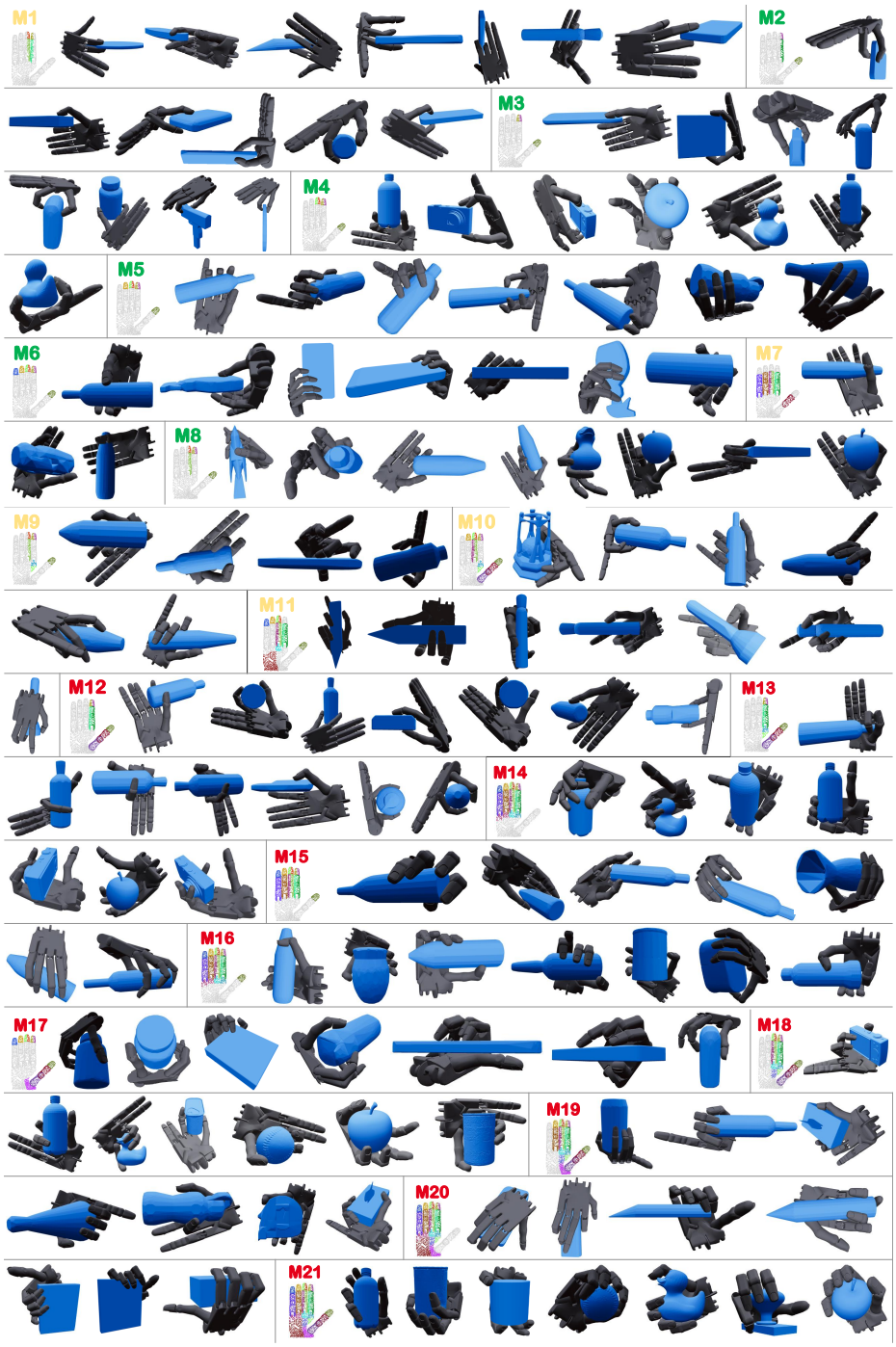

Fig. 13: Topological Clustering of Shadow Hand Grasps. Grasps grouped by contact topology (M1-M21), demonstrating consistent semantic alignment across varied object classes.