

# Generate, Transfer, Adapt: Learning Functional Dexterous Grasping from a Single Human Demonstration

Xingyi He<sup>1,2</sup>, Adhitya Polavaram<sup>2</sup>, Yunhao Cao<sup>2</sup>, Om Deshmukh<sup>2</sup>, Tianrui Wang<sup>2</sup>, Xiaowei Zhou<sup>1</sup>, Kuan Fang<sup>2†</sup>

**Abstract**—Functional grasping with dexterous robotic hands is a key capability for enabling tool use and complex manipulation, yet progress has been constrained by two persistent bottlenecks: the scarcity of large-scale datasets and the absence of integrated semantic and geometric reasoning in learned models. In this work, we present CorDex, a framework that robustly learns dexterous functional grasps of novel objects from synthetic data generated from just a single human demonstration. At the core of our approach is a correspondence-based data engine that generates diverse, high-quality training data in simulation. Based on the human demonstration, our data engine generates diverse object instances of the same category, transfers the expert grasp to the generated objects through correspondence estimation, and adapts the grasp through optimization. Building on the generated data, we introduce a multimodal prediction network that integrates visual and geometric information. By devising a local–global fusion module and an importance-aware sampling mechanism, we enable robust and computationally efficient prediction of functional dexterous grasps. Through extensive experiments across various object categories, we demonstrate that CorDex generalizes well to unseen object instances and significantly outperforms state-of-the-art baselines. For additional results and videos, please visit <https://cordex-manipulation.github.io>.

## I. INTRODUCTION

Functional grasping with dexterous hands is a fundamental capability that enables robots to perform complex tool use and precise manipulation. Unlike conventional grasping with simple end-effectors [1] or task-agnostic methods focused solely on stability [2], [3], dexterous functional grasping requires predicting high-dimensional motor commands that jointly satisfy both physical and semantic constraints [4], [5]. In particular, the robot must not only establish a stable hold on the object but also meaningfully interact with its task-relevant part in order to realize its intended functionality, as illustrated in Fig. 1. Satisfying these demands under contact-rich interactions and fine-grained control makes dexterous functional grasping a persistent challenge.

An increasing number of recent works have explored functional dexterous grasping with learning-based approaches. However, despite encouraging progress, advancement remains constrained by two fundamental bottlenecks. First, acquiring large-scale, high-quality datasets with functional

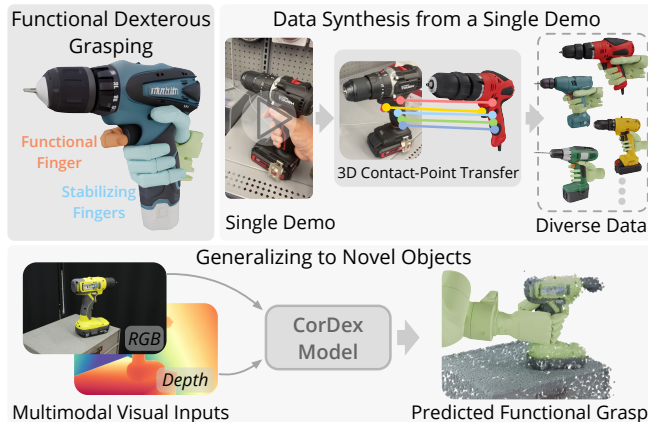


Fig. 1. CorDex learns to robustly perform functional dexterous grasping by combining a correspondence-based data engine and a multimodal grasp prediction model. The data engine scales a single human demonstration into diverse high-quality grasp data on novel objects. By learning from the generated data, the CorDex model leverages multimodal inputs to predict grasps for novel object instances.

dexterous grasp annotations is prohibitively difficult. Real-world data collection through motion capture or teleoperation [6]–[8] demands extensive human effort and scales poorly to novel objects and tasks. Alternatively, methods that leverage in-the-wild human video demonstrations [9], [10] offer broader coverage but suffer from severe reconstruction and pose estimation errors, necessitating costly data curation. Second, even with sufficient data, most approaches focus on geometric reasoning over object shape, which is inadequate for capturing semantic cues about functionality. Without jointly exploiting semantic and geometric information, these models often fail to produce grasps that are both physically stable and functionally appropriate in unseen scenarios.

In this paper, we present CorDex, a framework that enables learning robust dexterous functional grasping from a single human demonstration video. CorDex combines a data engine that autonomously produces diverse, high-quality training data in simulation with a novel prediction model that effectively integrates visual and geometric information to compute functional dexterous grasps for novel objects.

We devise a three-stage data engine to scale up functional dexterous grasping data from a single human video demonstration. First, given the object category in the video, the data engine generates diverse object instances by retrieving Internet images and converting them into 3D models. Second, the expert grasp from the demonstration is transferred to each generated object using a novel correspondence-based

<sup>1</sup>Xingyi He and Xiaowei Zhou are with Zhejiang University. This work was conducted while Xingyi He was a visiting scholar at Cornell University.

<sup>2</sup>Adhitya Polavaram, Yunhao Cao, Om Deshmukh, Tianrui Wang, and Kuan Fang are with Cornell University.

<sup>†</sup>Corresponding author.

pipeline. Finally, to ensure label quality, we introduce a physics-informed adaptation procedure that optimizes the transferred grasps in simulation. Unlike prior work that relies on 3D correspondence estimation [11], [12], which performs poorly due to the significant appearance and shape gap across different instances, our pipeline produces diverse and high-quality training data for functional dexterous grasping.

Building on the generated grasping data, we propose a prediction model, which learns to infer dexterous functional grasps from single-view RGB-D input. In contrast to previous approaches that rely solely on object geometry [3], [7], our model jointly reasons over semantic cues from images and geometric properties from point clouds. To achieve this, we design a local-global fusion module that integrates features from both modalities. In addition, we introduce a sampling mechanism that adaptively focuses on regions where robot-object interactions are likely to occur, improving both computational efficiency and prediction accuracy. Together, these components enable our model to produce grasps that are not only physically robust but also functionally meaningful for novel objects.

The proposed functional grasping data engine autonomously generates 11 million grasp-image pairs for 900 diverse objects across nine categories with minimal human annotation effort. Using both the generated diverse functional grasping data and the effective grasp prediction network, CorDex robustly generalizes to unseen objects from single-view input. We validate our approach through extensive experiments in simulation and the real world, spanning nine object categories and two robot embodiments. On unseen real-world objects, our method achieves a 69% success rate, substantially outperforming state-of-the-art methods.

## II. RELATED WORKS

**Dexterous grasping.** Grasping has long been a fundamental task in robotics. Early methods for dexterous hands planned grasps using analytic metrics such as force-closure and wrench-based quality measures [13]–[15]. With the rise of deep learning, data-driven approaches emerged that predict grasp configurations in different forms, including joint angles [16]–[18], contact regions [19], [20], and distance matrices between hand and object points [3]. These methods typically utilize geometry information derived from complete meshes [19], [20] or partial point clouds [3], [18]. To construct training data, grasp generation is often formulated as an optimization problem satisfying stability constraints [2], [21]. While effective for stable grasping, these geometry-centric approaches neglect semantic cues critical for functionality. Our work differs by jointly leveraging semantic features from RGB images and geometric features from point clouds, enabling predictions that are both stable and functional. Grasping has also been studied through category-level pose estimation, which predicts the 6D pose and 3D size of unseen objects from RGB-D data using category-level pretraining [22], [23], and transfers grasps from a reference object via the estimated transformation. However, because pose estimation provides only coarse alignment and

ignores fine-grained shape variations, the transferred grasps often miss the correct functional regions. In contrast, our approach directly predicts grasp gestures, enabling flexible generalization to unseen objects with large shape variations.

**Functional grasping and tool use.** Functional grasping requires contact with task-relevant object regions that afford intended use while ensuring stability. Classical approaches introduced task-oriented grasp metrics for dexterous hands [4], [24] and explored affordance reasoning or keypoint prediction for parallel-jaw grippers [25], [26]. More recent work has pursued learning-based solutions by collecting functional dexterous grasp data via motion capture [8], teleoperation [7], retargeting [27], or multimodal sensing [6]. However, these approaches require extensive human effort and scale poorly. To improve scalability, Internet demonstrations have been leveraged [9], [10], though reconstruction errors limit label quality. Some one-shot approaches generate functional grasps by transferring contact information across instances through dense 3D correspondences [11], [12], [28], [29]. However, these methods fall short due to limited generalizability to unseen objects, even within the same category, due to the limited 3D correspondence training data. In contrast, we propose a correspondence-based data engine that robustly transfers contact information across categories with a multimodal prediction model that fuses semantic and geometric information from RGB-D input to achieve robust prediction and category-level generalization to novel objects based on only a single human demonstration video for each object category.

**Robot learning from synthetic data.** Synthetic data has become a powerful enabler for scalable robot learning, reducing dependence on costly real-world annotation [30]–[33]. Prior works augment real-world demonstrations by generating diverse variants in simulation [34]–[36]. Recent advances in image matching [37]–[39] further enable scalable annotation transfer across object instances. Building on these ideas, our contribution targets the unique challenges of dexterous functional grasping. We introduce a correspondence-based engine that generates diverse, contact-rich grasps from a single human video demonstration.

## III. PRELIMINARIES

**Dexterous functional grasping.** Extending the problem formulation from [5], [9], [10], we consider a robotic hand with  $M$  fingers and  $K$  degrees of freedom. A dexterous grasp is defined as  $g = (T, \theta)$ , where  $T \in SE(3)$  is the hand pose and  $\theta \in \mathbb{R}^K$  represents finger joint angles. We require each grasp to satisfy two criteria: *functionality* and *stability*. A grasp is functionally appropriate if a designated set of fingers  $\mathcal{F} \subseteq \{1, \dots, M\}$  establishes contact with the corresponding functional regions  $\mathcal{R}_f \subseteq \mathbb{R}^3$  of the object:

$$\forall f \in \mathcal{F}, \exists p_f \in \mathcal{R}_f \text{ s.t. } \text{dist}(h_f(g), p_f) < \epsilon, \quad (1)$$

where  $h_f(g)$  is the fingertip position of finger  $f$  and  $\epsilon > 0$  is a tolerance. A grasp is defined as stable if a subset of stabilizing fingers  $\mathcal{S} \subseteq \{1, \dots, M\}$  can securely hold the object and resist external forces acting on the object, thereby

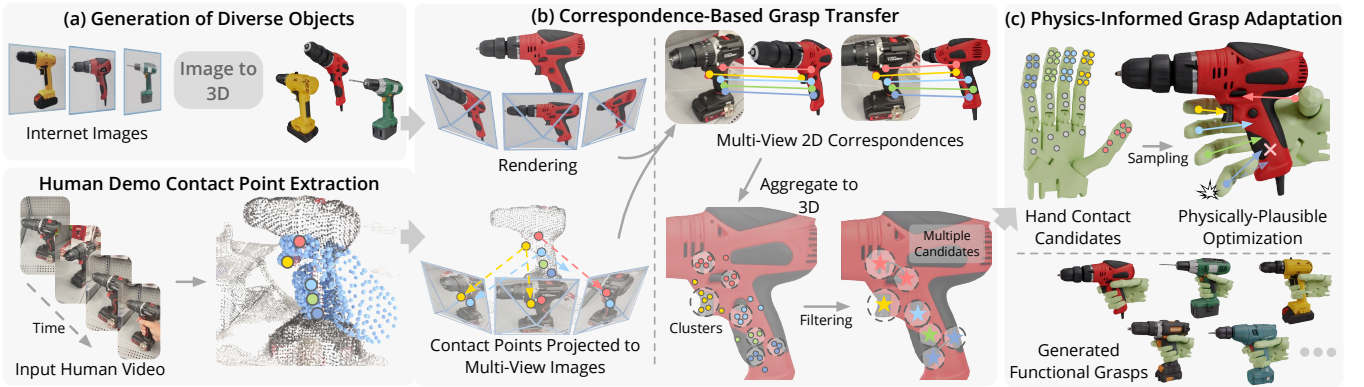


Fig. 2. **CorDex data engine.** We generate diverse, high-quality functional grasps for novel objects from a single human demonstration through three stages: (a) *Generate*: diversify objects within the task category by creating 3D models from Internet-retrieved images. (b) *Transfer*: extract 3D fingertip contacts (●●●●) from the demonstration via scene and hand reconstruction, then transfer them to novel objects using a correspondence-based 2D–3D pipeline that projects, matches, and aggregates contact points into reliable 3D candidates (★★) on generated objects. (c) *Adapt*: apply physics-informed grasp adaptation to convert candidate contact points into embodiment-specific grasps that satisfy both functionality and stability considerations, yielding diverse and high-quality grasp data.

maintaining its pose relative to the hand under perturbations. Together, these criteria capture the dual requirements that the grasp aligns with the object’s intended functionality while being robust to external disturbances. The robot receives a single-view RGB-D image and the target object mask, both aligned to the base frame through hand–eye calibration, and predicts functional grasps.

**Grasp prediction with  $\mathcal{D}(\mathcal{R}, \mathcal{O})$ .** To efficiently predict high-dimensional dexterous grasps, we build upon the  $\mathcal{D}(\mathcal{R}, \mathcal{O})$  paradigm [3], which uses a *dense distance matrix* between points sampled on the robot hand and object to represent a grasp. From this representation,  $g$  is recovered by multilateration [40] and inverse kinematics. Such a representation removes the need for expensive collision terms during optimization, and naturally generalizes across different hand embodiments.  $\mathcal{D}(\mathcal{R}, \mathcal{O})$  additionally formulates the policy network as a conditional variational autoencoder (CVAE), allowing the model to capture the multimodality of grasps contained in the training dataset. During training, the ground-truth hand configuration is encoded into a latent vector, which is concatenated with the fused object features to condition the distance matrix decoder. At test time, diverse grasps can be generated by discarding the latent encoder and directly sampling the latent from a prior distribution. The training objective combines three components: (i) an L1 loss between the predicted and ground-truth distance matrices, (ii) KL divergence regularization on the CVAE latent space, and (iii) a pose error loss supervising the reconstructed grasp. Building on the  $\mathcal{D}(\mathcal{R}, \mathcal{O})$  framework, our model adopts the dense distance matrix representation and the CVAE formulation, while introducing several key enhancements to enable accurate functional grasping.

#### IV. CORRESPONDENCE-BASED DATA SYNTHESIS

To enable scalable learning of functional dexterous grasping, we generate synthetic grasp data for each object category from a single human demonstration video. Such demonstrations can be easily captured with affordable devices such as smartphone cameras, avoiding the need for expensive multi-camera or teleoperation setups and allowing functional grasp

annotations with minimal manual effort. The central challenge, however, is how to scale from a single demonstration to diverse objects while ensuring that the synthesized grasp labels remain consistent with the demonstration video.

To address this challenge, we introduce a data engine that operates in three stages. First, we generate diverse object models to support grasp synthesis (Fig. 2a). Next, we extract 3D contact points from a single human demonstration to represent hand–object interactions, and transfer these points to novel objects through a robust cross-instance 2D–3D correspondence pipeline (Fig. 2b), which leverages advances in image matching and geometric cues to mitigate noise and inconsistencies. Finally, we apply a physics-informed grasp adaptation procedure that optimizes robot hand configurations with respect to the transferred contacts, ensuring both functional alignment and physical stability (Fig. 2c). The following subsections detail the design of each stage.

##### A. Generation of Diverse Objects

To enable category-level generalization, we generate diverse 3D object models that capture large intra-class variations while preserving the functional semantics of the task. Instead of relying on fixed 3D datasets with limited coverage or text-to-3D generation, where vague descriptions often lead to unrealistic shapes, we adopt a 2D-to-3D generation approach that leverages the broad visual diversity of Internet images. Starting from the demonstration video, we retrieve a large collection of Internet images of the same object category. Candidate images are filtered using pretrained visual feature similarity [37] to the demonstration, ensuring both diversity and relevance. In cases where the task category lacks sufficient Internet images, we augment them using GPT-Image [41] inpainting, and generate high-quality 3D object meshes with Rodin [42]. The retrieved images are then used as conditions for a 2D-to-3D generation model [42], producing massive, high-quality object meshes per object category. Compared to using a fixed object dataset, this approach creates realistic and diverse assets tailored to the demonstration, providing a strong foundation for cross-instance functional grasp transfer.

## B. Correspondence-Based Grasp Transfer

With diverse object models generated for each task category, the next step is to transfer functional grasp knowledge from the human demonstration onto these novel instances. Directly retargeting human hand poses to a robot is infeasible due to object misalignment and the morphology gaps between human and robot hands. Instead, we represent human-object interaction through *3D fingertip contact keypoints*, which are embodiment-agnostic and transferable across objects. From the demonstration video, we reconstruct the hand mesh [43] and the object point cloud [44], and then extract fingertip contacts as the nearest object surface points. Since the reconstructed point clouds lack absolute scale, we determine the optimal scale by aligning the object to the hand mesh and minimizing the distances between fingertip points and their nearest object points.

Transferring these contact points to diversified objects is challenging because of large appearance and geometry variations. Naively applying cross-instance 3D matching [11], [12] performs poorly (Sec. VI) due to limited training data and weak generalization. To overcome this, we leverage large-scale pretrained 2D matching models [37], [39], which generalize across categories, and couple them with a robust 3D aggregation step. Specifically, the fingertip contacts are projected onto all valid frames of the demonstration, while novel objects are rendered from viewpoints uniformly sampled on a sphere. A 2D matcher [39] establishes correspondences between demonstration frames and rendered images, enabling contact points to be transferred to novel object renderings and subsequently back-projected into 3D using the known camera intrinsics, extrinsics, and depth.

Because 2D matching can be noisy and view-inconsistent, the back-projected points from multiple views are aggregated in 3D with density-based clustering. We retain the centers of the three largest clusters as candidate contact locations for each fingertip and discard smaller clusters as outliers. To further improve reliability, each candidate is weighted by the average 2D matching confidence of its member points. The resulting candidate set provides multiple plausible, confidence-weighted hypotheses for each fingertip, which are then resolved through the physics-informed grasp optimization. Preserving this set of hypotheses explicitly models cross-instance ambiguity and allows downstream optimization to exploit geometric and physical constraints to select stable, functional grasps.

## C. Physics-Informed Grasp Adaptation

Based on the candidate contact points on novel objects that specify the target locations for each finger, our goal is to generate embodiment-specific grasp labels for downstream model training. However, variations in object scale between the demonstration and generated objects, together with correspondence noise, may make the transferred contact points unreachable for the robot hand. To address this, we introduce a grasp adaptation process that jointly optimizes contact-point alignment and physical plausibility, ensuring that the resulting grasps are both functional and stable.

Specifically, for a robot hand, a set of candidate contact points is defined on every finger, which are likely to correspond to the transferred contact points on the object. Considering size variations between the robot and human hands, we define candidate contact points on both the middle and distal links to provide more flexibility in aligning with the transferred contacts, as shown in Fig. 2c. Our pipeline simultaneously initializes  $N$  grasps  $g$  and optimizes them with hand contact points and object contact points sampled from the candidates. The optimization objective is composed of the following terms:

- *Contact-prior loss* encourages sampled points on the hand surface to align with sampled contact points on the object surface, minimizing both positional distance and the deviation between normals:

$$\mathcal{L}_{prior} = \sum_{l \in \mathcal{C}} (\|h_l(g) - o_p\|_2^2 + \alpha(1 - n_h^\top n_o)), \quad (2)$$

where  $\mathcal{C}$  denotes the set of finger links with defined contact points,  $h_l(g)$  is the 3D position of a sampled finger contact point on link  $l$  under grasp  $g$ ,  $o_p$  is the transferred prior contact point on the object surface,  $n_h$  and  $n_o$  are their surface normals, and  $\alpha$  is a hyperparameter balancing positional and orientation terms.

- *Stability contact loss* addresses scale mismatch (transferred contact points may be too dense or sparse to be reachable). This loss aligns sampled finger points (the same points as those used in the *contact-prior loss*) with their nearest object surface samples, reducing unstable floating gestures caused by object scale misalignment:

$$\mathcal{L}_{stab} = \sum_{l \in \mathcal{C}} \|h_l(g) - o_c\|_2^2, \quad (3)$$

where  $h_l(g)$  is a sampled finger contact point under hand configuration  $g$ , and  $o_c$  denotes its closest neighbor on the object surface.

- *Auxiliary contact loss*. Since the transferred points constrain only the middle and distal finger links, we additionally sample points on other hand links (e.g., the palm) and encourage them to contact the object surface, thereby improving overall stability:

$$\mathcal{L}_{aux} = \sum_{l \in \mathcal{A}} \|h_l(g) - o_c\|_2^2, \quad (4)$$

where  $\mathcal{A}$  is the set of auxiliary hand links,  $h_l(g)$  is a sampled contact point on link  $l$ , and  $o_c$  denotes its closest neighbor on the object surface.

- *Joint limit loss* penalizes violations of joint angle limits to ensure feasible configurations.
- *Collision loss* penalizes the robot-object penetration.
- *Self-penetration loss* penalizes robot links' penetration.

The formulations for *joint-limit*, *collision*, and *self-penetration losses* follow [2]. The final objective is the weighted sum of all loss terms and is optimized using the method in [45]. After optimization, we obtain candidate grasps, which are verified in simulation to ensure physical stability [2]. Specifically, each grasp is tested in a real-time physics engine [46] by applying external forces from six

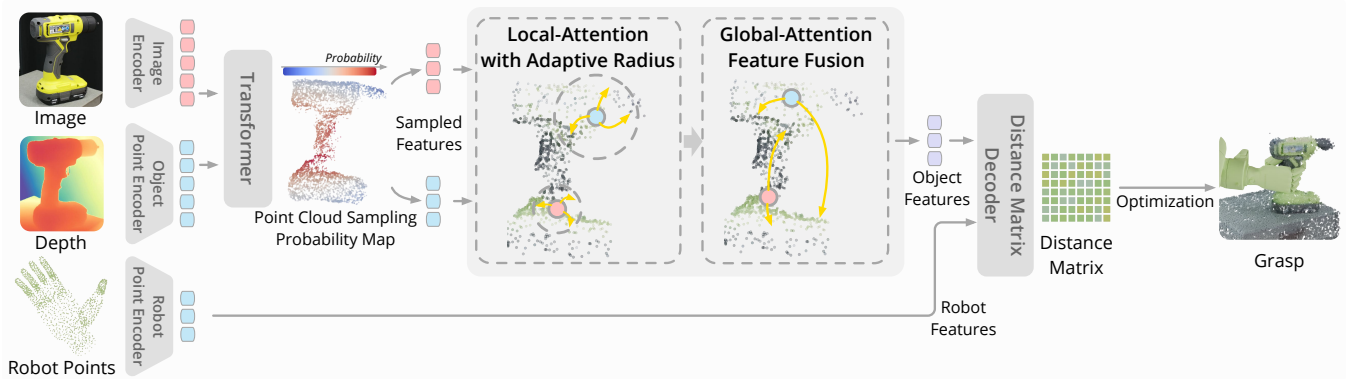


Fig. 3. **CorDex grasp prediction network.** The network integrates semantic and geometric information from single-view RGB-D input to predict functional dexterous grasps for novel objects. Image and point cloud features are first encoded into pointwise features and processed by a transformer. To boost performance and computational efficiency, we introduce an importance-aware sampling mechanism that samples points around contact areas. Given the sampled points, a local-global fusion module refines local details and encodes holistic object context through global attention. Finally, a distance matrix between the robot hand and object points is decoded via cross-attention and optimized to obtain the final grasp.

directions to check whether the object remains securely held. Verified grasps are retained as functional grasping data for training the prediction network. Finally, to produce large-scale training data for grasp prediction, we render photo-realistic RGB-D images in Blender [47], placing objects with random poses within a  $1\text{m}^3$  cube and enhancing diversity with randomized backgrounds and lighting.

## V. GRASP PREDICTION VIA MULTIMODAL FUSION

We propose a novel functional dexterous grasp prediction model trained on the large-scale dataset generated by our data pipeline. The model builds on the  $\mathcal{D}(\mathcal{R}, \mathcal{O})$  representation [3] introduced in Sec. III, where a grasp is encoded as a dense distance matrix between sampled points on the robot hand and the object, and the final grasp  $g$  is recovered through multilateration and inverse kinematics. Unlike  $\mathcal{D}(\mathcal{R}, \mathcal{O})$ , which relies exclusively on point cloud geometry, our approach jointly exploits semantic cues from RGB images and geometric properties from depth observations. This multimodal integration enables the model to predict grasps that simultaneously satisfy *stability* and *functionality*, both of which are essential for dexterous functional grasping.

As illustrated in Fig. 3, the network takes as input a single-view RGB-D observation of the object together with the robot hand point cloud. The depth channel is converted into a 3D point cloud, while semantic features are extracted from the RGB image and back-projected onto the same 3D points. Each point is thus represented by both geometric and semantic embeddings computed with image and point cloud encoders [37], [48]. These multimodal pointwise features are fused into a unified object representation, which is cross-attended with the robot hand features to predict the distance matrix. Two design challenges motivate our architecture: (i) functional regions such as triggers or buttons are often small and easily missed by uniform point sampling, and (ii) accurate grasping requires reasoning at both local detail and global object context. To address these challenges, we introduce two central components: an *importance-aware sampling module* that adaptively focuses on contact-relevant regions, and a *local-global fusion module* that integrates

complementary semantic and geometric cues into a coherent representation.

**Local-global fusion of multimodal features.** Functional dexterous grasping requires reasoning over both the fine-grained geometry of contact regions and the broader semantic context of the object. To address this, we design a local-global fusion module that adaptively combines semantic and geometric information with different receptive fields. Specifically, semantic features from the RGB image are back-projected onto the sampled 3D points and paired with their geometric features. A local cross-attention mechanism then relates geometric features to nearby semantic features, and vice versa, enabling the network to capture contact-relevant detail. To account for varying point densities, we introduce an *adaptive attention radius*, which uses a larger receptive field to gather richer context in sparse regions, while focusing more narrowly for precision in dense regions. The resulting locally fused features are further processed with global self-attention to encode the object’s overall structure. Finally, local and global features are integrated into a unified representation, which is cross-attended with robot hand features to predict the distance matrix as described in Sec. III.

**Efficient prediction via adaptive sampling.** Encoding all object points uniformly is both computationally inefficient and ineffective for functional grasping, since small functional regions (*e.g.*, triggers, buttons) may be overwhelmed by irrelevant surface points. To focus computation on task-relevant areas, we introduce an importance-aware sampling module that adaptively preserves points likely to lie in contact regions. Given the concatenated semantic and geometric features of each object point, a lightweight transformer estimates pointwise importance probabilities using global self-attention to incorporate object context. Guided by this distribution, the point cloud is downsampled from  $N = 4096$  to  $N' = 1024$  points, increasing density in functional regions while reducing redundancy elsewhere. Ground-truth importance maps are derived from distances between object points and ground-truth robot hand points, and the sampling module is trained with the KL divergence to match this distribution.

TABLE I

**QUANTITATIVE COMPARISONS IN SIMULATION.** WE EVALUATE OUR METHOD AGAINST STATE-OF-THE-ART APPROACHES ON ALL NINE TASKS USING TWO ROBOTIC HANDS: THE 22-DOF SHADOW HAND AND THE 6-DOF INSPIRE HAND. REPORTED SUCCESS RATES (%) INDICATE GRASPS THAT SATISFY BOTH STABILITY AND FUNCTIONALITY. THE BEST RESULTS ARE SHOWN IN **BOLD**. \* DENOTES ONE-SHOT METHODS

| Embodiment | Method  | Drill       | Pipette     | Stapler     | Spray Bottle | Hammer      | Syringe     | Hair Dryer  | Aerosol Can | Glue Gun    | Avg.        |
|------------|---|-------------|-------------|-------------|--------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Shadow     | $\mathcal{D}(\mathcal{R}, \mathcal{O})$ [3]               | 24.0        | 11.7        | 23.3        | 19.0         | 14.3        | 28.0        | 8.7         | 25.3        | 10.0        | 18.3        |
|            | $\mathcal{D}(\mathcal{R}, \mathcal{O})$ [3] with our data | 37.7        | 20.7        | 33.7        | 37.7         | 21.0        | 48.0        | 70.0        | 33.3        | 21.7        | 36.0        |
|            | SparseDFF* [11]   | 7.7         | 14.7        | 15.7        | 16.3         | 22.0        | 18.7        | 11.3        | 17.0        | 9.7         | 14.8        |
|            | DenseMatcher* [12]  | 14.3        | 16.7        | 15.3        | 19.7         | 25.3        | 16.3        | 18.3        | 15.0        | 11.0        | 16.9        |
|            | AG-Pose [23] with our data                                | 67.7        | 65.3        | 76.0        | 63.3         | 77.0        | 71.3        | 69.0        | 59.0        | 58.7        | 67.5        |
|            | <b>Ours</b>   | <b>90.0</b> | <b>91.3</b> | <b>85.0</b> | <b>85.3</b>  | <b>85.7</b> | <b>91.7</b> | <b>98.7</b> | <b>84.3</b> | <b>84.7</b> | <b>88.5</b> |
| Inspire    | $\mathcal{D}(\mathcal{R}, \mathcal{O})$ [3] with our data | 13.3        | 11.7        | 17.0        | 26.3         | 18.0        | 13.7        | 25.0        | 7.7         | 25.3        | 17.6        |
|            | SparseDFF* [11]   | 8.0         | 6.0         | 10.3        | 2.3          | 12.3        | 7.3         | 9.7         | 6.3         | 7.7         | 7.8         |
|            | DenseMatcher* [12]  | 5.3         | 6.3         | 7.0         | 4.7          | 14.3        | 7.7         | 12.3        | 7.0         | 3.7         | 7.6         |
|            | AG-Pose [23] with our data                                | 41.3        | 47.0        | 60.7        | 56.0         | 58.3        | 49.0        | 40.0        | 44.3        | 43.3        | 48.9        |
|            |   | <b>Ours</b> | <b>72.7</b> | <b>63.3</b> | <b>80.3</b>  | <b>87.7</b> | <b>78.0</b> | <b>73.0</b> | <b>75.3</b> | <b>70.7</b> | <b>71.0</b> |



Fig. 4. **Examples of generated data.** We generate a functional dexterous grasp dataset consisting of 900 objects, 1.08 million images, and 11 million image–grasp pairs. The dataset spans across nine tasks and two different embodiments of different DoFs (Shadow and Inspire).

This adaptive sampling not only improves computational efficiency but also enhances the model’s ability to capture the subtle object regions essential for functional grasp prediction.

## VI. EXPERIMENTS

We conduct extensive experiments in both simulation and the real world to evaluate the effectiveness of our approach.

### A. Experimental Setup

**Evaluation environments and protocols.** We consider both the stability and functionality of predicted grasps using simulation and real-world experiments, reporting the success rate of grasps satisfying both requirements.

*Simulation.* We conduct experiments across nine object categories using a held-out test set with object models, annotated functional regions, and avoidance regions that should not be touched by the hand. Generated grasps are validated in IsaacGym [46] under external forces on two embodiments: the 22-DoF Shadow Hand and the 6-DoF Inspire Hand, following the protocol in [3]. A grasp is considered stable if object displacement is  $< 2$  cm after external forces are applied, while it is considered functional if the distance between robot hand and functional region  $< 1$  mm and no avoidance regions (e.g. drill head) are touched.

*Real-world.* We evaluate across six object categories, each containing 3 objects, using the 6-DoF OYMotion hand mounted on a 7-DoF Franka Research 3 arm, as shown in Fig. 5. The OYMotion hand is nearly identical to the Inspire Hand but is produced by a different manufacturer. A ZED camera is mounted on either side of the table and provides single-view RGB-D input. We use Grounded SAM [49], [50]

to segment the target object. We evaluate on five poses of each object.

**Baseline methods.** Our method is compared with three categories of approaches: (1) Dexterous hand grasp prediction method:  $\mathcal{D}(\mathcal{R}, \mathcal{O})$  [3], which directly predicts grasps. We also report the results of  $\mathcal{D}(\mathcal{R}, \mathcal{O})$  trained on our dataset. (2) One-shot correspondence methods: SparseDFF [11] and DenseMatcher [12]. Since these methods cannot handle single-view input, we provide two-view RGB-D images in real-world experiments (captured from different viewpoints for a more complete observation) and complete object models in simulation. (3) Category-level object pose estimation method: AG-Pose [23], which is trained on our dataset for each category to ensure fair comparison.

### B. CorDex Dataset

Our data engine generates high-quality functional grasp data for diverse objects covering nine common functional object categories, enabling the grasp prediction model to robustly generalize to unseen objects. The object categories include: *Drill*, *Pipette*, *Stapler*, *Spray Bottle*, *Hammer*, *Syringe*, *Hair Dryer*, *Aerosol Can*, and *Glue Gun*, as shown in Fig. 4. For each category, we create 100 diverse objects with varying shapes and appearances but consistent functionality. For each object, 10 valid functional grasps are generated for both the Shadow Hand and Inspire Hand, two widely used robotic embodiments. We render 1,200 RGB-D images per object under diverse poses and lighting conditions. In total, the dataset contains 900 objects, 1.08 million images, and around 11 million image–grasp pairs, generated in  $\sim 3$  days using 48 NVIDIA A100 GPUs. For each task, 2 objects are held out for validation and 3 for testing, ensuring that evaluation is performed on unseen instances.

### C. Comparative Results

The quantitative results of simulation and real-world experiments are reported in Tab. I and Tab. II, while qualitative results are shown in Fig. 5. Compared with *grasp prediction method*  $\mathcal{D}(\mathcal{R}, \mathcal{O})$  [3], our method achieves significantly higher performance in both simulation and real-world settings.  $\mathcal{D}(\mathcal{R}, \mathcal{O})$  performs poorly due to the lack of functional grasp supervision in its original training data; retraining it on our dataset improves results but still performs substantially

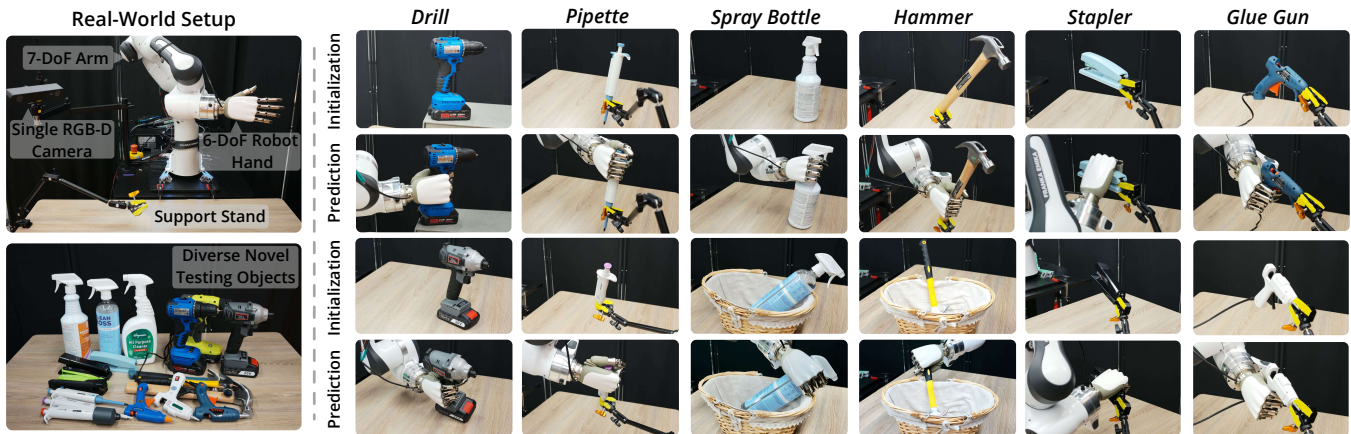


Fig. 5. **Real-world experiments.** Left: a 7-DoF robot arm with a 6-DoF dexterous hand executes functional grasps predicted by our model from single-view RGB-D input. We evaluate six tasks, each with three real-world objects that are unseen in the generated dataset. Right: qualitative results on these objects, demonstrating category-level generalization to diverse shapes and varying poses.

TABLE II

**QUANTITATIVE COMPARISON IN THE REAL-WORLD. SUCCESS COUNTS FOR FUNCTIONAL GRASPING ACROSS SIX TASKS. BEST RESULTS ARE SHOWN IN **BOLD**. \* DENOTES ONE-SHOT METHODS.**

|                                 | Drill        | Pipette     | Stapler      | Spray Bottle | Hammer       | Glue Gun     |
|---------------------------------|--------------|-------------|--------------|--------------|--------------|--------------|
| <i>D(R,O)</i> [3] with our data | 2/15         | 0/15        | 3/15         | 2/15         | 4/15         | 2/15         |
| SparseDFF* [11]                 | 3/15         | 0/15        | 3/15         | 1/15         | 3/15         | 1/15         |
| DenseMatcher* [12]              | 1/15         | 0/15        | 2/15         | 0/15         | 3/15         | 0/15         |
| AG-Pose [23] with our data      | 3/15         | 2/15        | 6/15         | 3/15         | 9/15         | 4/15         |
| <b>Ours</b>                     | <b>10/15</b> | <b>7/15</b> | <b>11/15</b> | <b>11/15</b> | <b>13/15</b> | <b>10/15</b> |

worse than our approach, highlighting the effectiveness of our model design in utilizing visual and features. Compared with *one-shot correspondence* methods SparseDFF [11] and DenseMatcher [12], which are provided with more complete observations, our single-view method also outperforms them by a large margin. These methods suffer from the lack of generalizability of dense 3D correspondence, which is difficult to learn from limited correspondence data, leading to inaccurate contact transfer and low-quality grasps. Finally, compared with the *category-level pose estimation* method AG-Pose [23], our approach achieves substantially better results. Even when trained on our dataset, category-level methods rely on coarse object alignment, which is insufficient for functional grasping that requires precise contact with functional regions. In contrast, our method directly predicts grasp gestures without explicit alignment, demonstrating generalization to unseen objects with diverse shapes.

#### D. Ablation Studies

We conduct ablation studies on six tasks in simulation with the Inspire Hand to evaluate the contribution of each design component in both the data generation pipeline and the model. For the *data generation*, we first replace our correspondence transfer with a 3D matching method [12]. As shown in Tab. III(1), this substitution leads to a significant performance drop due to inaccurate contact transfer and low-quality grasp optimization, highlighting the effectiveness and flexibility of our robust correspondence transfer approach. We further ablate the design of preserving multiple transferred contact points during optimization (Tab. III(2)). Retaining only a single contact point results in degraded

TABLE III

**ABLATION STUDIES IN SIMULATION. WE REPORT SUCCESS RATE (%) FOR EACH VARIANT. BEST RESULTS ARE SHOWN IN **BOLD**.**

| Method   | Drill       | Pipette     | Stapler     | Hammer      | Aerosol Can | Glue Gun    | Avg.        |
|--|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Full   | <b>72.7</b> | <b>63.3</b> | <b>80.3</b> | <b>78.0</b> | <b>70.7</b> | <b>71.0</b> | <b>72.7</b> |
| (1) Data engine with 3D matching [12]            | 25.3        | 14.0        | 16.3        | 11.7        | 18.7        | 25.0        | 18.5        |
| (2) Data engine <i>w/o</i> multiple candidates   | 67.3        | 59.7        | 71.0        | 71.3        | 62.0        | 65.3        | 66.1        |
| (3) Grasp network <i>w/o</i> image input         | 20.0        | 18.0        | 23.3        | 15.0        | 19.7        | 28.0        | 20.7        |
| (4) Grasp network <i>w/o</i> importance sampling | 64.7        | 57.0        | 73.3        | 66.3        | 60.7        | 68.3        | 65.1        |
| (5) Grasp network <i>w/o</i> local attention     | 47.3        | 51.7        | 60.0        | 57.0        | 51.3        | 49.0        | 52.7        |

performance, confirming the importance of multiple candidates for handling transfer noise. For the *model architecture*, we first remove the image input and use only point clouds (Tab. III(3)). This causes a large performance drop, as single-view point clouds provide ambiguous semantics, while our model benefits from complementary image features. Next, we ablate the importance-aware sampling (Tab. III(4)). Without it, fewer points near the hand are preserved and performance decreases noticeably. Finally, removing the local attention module with adaptive radius (Tab. III(5)) consistently lowers performance, demonstrating its effectiveness in capturing fine-grained local context as a complement to global attention.

## VII. CONCLUSION

In this paper, we present a novel framework for dexterous functional grasping that integrates a correspondence-based data engine with a grasp prediction network employing local-global adaptive feature fusion. The data engine autonomously generates large-scale functional grasp data for diverse objects from a single human demonstration, while the prediction network effectively leverages both visual and geometric features to infer accurate functional grasps from single-view RGB-D input. Extensive experiments in both simulation and real-world settings show that our approach substantially outperforms state-of-the-art baselines. Furthermore, our data engine can be readily extended to new tasks without additional training, offering a scalable pipeline for data curation and paving the way toward universal dexterous grasping models.

**Limitations.** Despite these advances, there are two main limitations of the proposed approach. First, although depth noise is injected during training, the model remains sensitive to

severely corrupted or displaced depth input in the real world, reflecting the domain gap between synthetic and real-world depth sensing. Second, while the framework generalizes to novel object instances, it still focuses on category-specific training and does not yet handle fully open-set scenarios. Future work should explore scaling task diversity and developing universal models that exhibit emergent generalization to unseen objects and tasks.

## REFERENCES

- [1] H. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu, "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains," *IEEE Transactions on Robotics*, 2022.
- [2] R. Wang, J. Zhang, J. Chen, Y. Xu, P. Li, T. Liu, and H. Wang, "Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation," *ICRA*, 2022.
- [3] Z. Wei, Z. Xu, J. Guo, Y. Hou, C. Gao, Z. Cai, J. Luo, and L. Shao, "D(r, o) grasp: A unified representation of robot and object interaction for cross-embodiment dexterous grasping," *ICRA*, 2025.
- [4] Z. Li and S. Sastry, "Task-oriented optimal grasping by multifingered robot hands," *IEEE J. Robotics Autom.*, 1987.
- [5] A. Agarwal, S. Uppal, K. Shaw, and D. Pathak, "Dexterous functional grasping," in *CoRL*, 2023.
- [6] S. Brahmabhatt, C. Ham, C. C. Kemp, and J. Hays, "Contactdb: Analyzing and predicting grasp contact via thermal imaging," *CVPR*, 2019.
- [7] Y. Liu, Y. Yang, Y. Wang, X. Wu, J. Wang, Y. Yao, S. Schwertfeger, S. Yang, W. Wang, J. Yu, et al., "Realdex: Towards human-like grasping for robotic dexterous hand," *IJCAI*, 2024.
- [8] L. Yang, K. Li, X. Zhan, F. Wu, A. Xu, L. Liu, and C. Lu, "Oakink: A large-scale knowledge repository for understanding hand-object interaction," *CVPR*, 2022.
- [9] A. Kannan, K. Shaw, S. Bahl, P. Mannam, and D. Pathak, "Defit: Dexterous fine-tuning for hand policies," in *CoRL*, 2023.
- [10] H. Chen, Y. Yao, Y. Ye, Z. Xu, H. Bharadhwaj, J. Wang, S. Tulsiani, Z. Erickson, and J. Ichnowski, "Web2grasp: Learning functional grasps from web images of hand-object interactions," *ArXiv*, vol. abs/2505.05517, 2025.
- [11] Q. Wang, H. Zhang, C. Deng, Y. You, H. Dong, Y. Zhu, and L. J. Guibas, "Sparsedff: Sparse-view feature distillation for one-shot dexterous manipulation," *ICLR*, 2023.
- [12] J. Zhu, Y. Ju, J. Zhang, M. Wang, Z. Yuan, K. Hu, and H. Xu, "Densematcher: Learning 3d semantic correspondence for category-level manipulation from a single demo," *ICLR*, 2024.
- [13] C. Ferrari and J. F. Canny, "Planning optimal grasps," *ICRA*, 1992.
- [14] C. W. Borst, M. Fischer, and G. Hirzinger, "Grasp planning: how to choose a suitable task wrench space," *ICRA*, 2004.
- [15] Y. Lin and Y. Sun, "Grasp planning to maximize task coverage," *IJRR*, 2015.
- [16] Y. Xu et al., "Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy," *CVPR*, 2023.
- [17] H. Geng and Y. Liu, "Unidexgrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning," *ICCV*, 2023.
- [18] H. Zhang, Z. Wu, L. Huang, S. Christen, and J. Song, "RobustDex-Grasp: Robust dexterous grasping of general objects," in *CoRL*, 2025.
- [19] L. Shao, F. Ferreira, M. Jorda, V. Nambiar, J. Luo, E. Solowjow, J. A. Ojea, O. Khatib, and J. Bohg, "Unigrasp: Learning a unified model to grasp with multifingered robotic hands," *RA-L*, 2019.
- [20] P. Li, T. Liu, Y. Li, Y. Geng, Y. Zhu, Y. Yang, and S. Huang, "Gendexgrasp: Generalizable dexterous grasping," *ICRA*, 2022.
- [21] J. Chen, Y. Ke, L. Peng, and H. Wang, "Dexonomy: Synthesizing all dexterous grasp types in a grasp taxonomy," *RSS*, 2025.
- [22] H. Wang, S. Sridhar, J. Huang, J. P. C. Valentin, S. Song, and L. J. Guibas, "Normalized object coordinate space for category-level 6d object pose and size estimation," *CVPR*, 2019.
- [23] X. Lin, W. Yang, Y. Gao, and T. Zhang, "Instance-adaptive and geometric-aware keypoint learning for category-level 6d object pose estimation," *CVPR*, 2024.
- [24] R. Haschke, J. J. Steil, I. Steuwer, and H. J. Ritter, "Task-oriented quality measures for dextrous grasping," *International Symposium on Computational Intelligence in Robotics and Automation*, 2005.
- [25] Y. Zhu, Y. Zhao, and S.-C. Zhu, "Understanding tools: Task-oriented object modeling, learning and recognition," *CVPR*, 2015.
- [26] K. Fang, Y. Zhu, A. Garg, A. Kurenkov, V. Mehta, L. Fei-Fei, and S. Savarese, "Learning task-oriented grasping for tool manipulation from simulated self-supervision," *IJRR*, 2018.
- [27] L. Huang, H. Zhang, Z. Wu, S. J. Christen, and J. Song, "Fungrasp: Functional grasping for diverse dexterous hands," *RA-L*, 2025.
- [28] R. Wu, T. Zhu, W. Peng, J. Hang, and Y. Sun, "Functional grasp transfer across a category of objects from only one labeled instance," *RA-L*, 2023.
- [29] W. Wei, P. Wang, S. Wang, Y. Luo, W. Li, D. Li, Y. Huang, and H. Duan, "Learning human-like functional grasping for multifinger hands from few demonstrations," *IEEE Transactions on Robotics*, 2024.
- [30] K. Fang, T. Migimatsu, A. Mandlekar, L. Fei-Fei, and J. Bohg, "Active task randomization: Learning robust skills via unsupervised generation of diverse and feasible tasks," in *IROS*. IEEE, 2023.
- [31] Y. Mu, T. Chen, Z. Chen, S. Peng, Z. Lan, Z. Gao, Z. Liang, Q. Yu, Y. Zou, M. Xu, L. Lin, Z. Xie, M. Ding, and P. Luo, "Robotwin: Dual-arm robot benchmark with generative digital twins," in *CVPR*, June 2025, pp. 27 649–27 660.
- [32] T. Lin, Z.-H. Yin, H. Qi, P. Abbeel, and J. Malik, "Twisting lids off with two hands," *arXiv:2403.02338*, 2024.
- [33] R. Wang, J. Zhang, J. Chen, Y. Xu, P. Li, T. Liu, and H. Wang, "Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation," 2023. [Online]. Available: <https://arxiv.org/abs/2210.02697>
- [34] Z. Jiang, Y. Xie, K. Lin, Z. Xu, W. Wan, A. Mandlekar, L. Fan, and Y. Zhu, "Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning," in *ICRA*, 2025.
- [35] T. Dai, J. Wong, Y. Jiang, C. Wang, C. Gokmen, R. Zhang, J. Wu, and L. Fei-Fei, "Automated creation of digital cousins for robust policy learning," in *CoRL*, 2024.
- [36] A. Maddukuri et al., "Sim-and-real co-training: A simple recipe for vision-based robotic manipulation," in *RSS*, Los Angeles, CA, USA, 2025.
- [37] M. Oquab et al., "Dinov2: Learning robust visual features without supervision," *ArXiv*, 2023.
- [38] Y. Wang, X. He, S. Peng, D. Tan, and X. Zhou, "Efficient LoFTR: Semi-dense local feature matching with sparse-like speed," in *CVPR*, 2024.
- [39] X. He, H. Yu, S. Peng, D. Tan, Z. Shen, H. Bao, and X. Zhou, "Matchanything: Universal cross-modality image matching with large-scale pre-training," in *Arxiv*, 2025.
- [40] A. Norrdine, "An algebraic solution to the multilateration problem," in *Proceedings of the 15th international conference on indoor positioning and indoor navigation, Sydney, Australia*, 2012.
- [41] OpenAI, "Addendum to gpt-4o system card: Native image generation," OpenAI, Tech. Rep., 2025.
- [42] D. T. Inc. Rodin ai - free ai 3d model generator for everyone. <https://hyper3d.ai/>.
- [43] R. A. Potamias, J. Zhang, J. Deng, and S. Zafeiriou, "Wilor: End-to-end 3d hand localization and reconstruction in-the-wild," *CVPR*, 2025.
- [44] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotný, "Vggg: Visual geometry grounded transformer," *CVPR*, 2025.
- [45] T. Liu, Z. Liu, Z. Jiao, Y. Zhu, and S.-C. Zhu, "Synthesizing diverse and physically stable grasps with arbitrary hand structures using differentiable force closure estimator," *RA-L*, 2021.
- [46] V. Makovychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, and G. State, "Isaac gym: High performance gpu-based physics simulation for robot learning," *ArXiv*, 2021.
- [47] Blender. [Online]. Available: <https://www.blender.org/>
- [48] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *TOG*, 2018.
- [49] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, et al., "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *Arxiv*, 2023.
- [50] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," *ICCV*, 2023.