Web2Grasp: Learning Functional Grasp with Reconstructed HOI from Web Images

Hongyi Chen[1*], Yunchao Yao[1*], Yufei Ye[2], Zhixuan Xu[3], Homanga Bharadhwaj[1],
Jiashun Wang[1], Shubham Tulsiani[1], Zackory Erickson[1] and Jeffrey Ichnowski[1]

Abstract— **Dexterous and functionally valid grasping is essential for enabling multi-finger robotic hands to manipulate objects effectively. However, most prior work either focuses solely on power grasping, which simply involves holding an object still, or relies on costly human-collected demonstrations to teach robots how to grasp each object functionally. Instead, we propose extracting human grasp information from internet images which depict natural and functional object interactions, thereby bypassing the need for curated human demonstrations. Leveraging existing 3D reconstruction methods from RGB images, we reconstruct hand-object interaction (HOI) meshes, retarget the human hand to the ShadowHand robot, and align the noisy object mesh with its accurate 3D shape. We demonstrate that low-quality HOI data from web sources can effectively train a functional grasping model. To further expand the grasp dataset for seen and unseen objects, we utilize the initially-trained grasping model with web data in IsaacGym simulator to generate physically feasible grasps while preserving functionality. The simulator-augmented dataset boosts the model's success rate from 61.8% to 83.4%. We train the model on 10 object categories and evaluate it on 9 unseen objects, including challenging items such as syringes, pens, spray bottles, and tongs, which are underrepresented in existing datasets. Evaluation results in IsaacGym show a 6.7% improvement in grasp success rates on unseen objects, and significantly higher functionality ratings from human evaluators. Project website is available at: `https://web2grasp.github.io/`.**

I. Introduction

Functional grasping with dexterous robot hands, which requires context-aware hand-object interactions, remains a significant challenge in multi-finger manipulation. While prior work has made substantial progress in dexterous grasping, many studies focus on generic power grasps, which often apply nearly identical grasps to objects that require diverse poses for effective operation [1–4]. These methods commonly rely on force-closure estimation or reinforcement learning approaches [5–7]. While learning functional affordances from images and object shapes has been explored [8–10], studies have shown that learning high-dimensional grasp poses from such noisy data remains challenging and often requires additional human demonstrations to fine-tune the trained prior [9, 11, 12]. High-quality, yet expensive, human demonstration datasets that account for the complexities of diverse object geometries and functional constraints have been developed [13–15], and these datasets are used to train robots for functional grasping [16–18].

To avoid the time-consuming human annotation of grasps across object categories, we turn to recent advances in hand-object pose estimation and interaction reconstruction from RGB images [19–21]. These methods can reconstruct human-object interactions (HOI) in the form of 3D representations [21–23], from images depicting how humans grasp objects in functional ways. However, due to mutual occlusion and the lack of annotated data, these existing models are often not accurate, with reconstructed HOI exhibiting excessive penetration or failing to make proper contact with object meshes [21, 24]. This inaccuracy has hindered the use of such data for training functional dexterous grasps. Some prior works have attempted to replay reconstructed 3D data in simulators [12, 25, 26] to improve the robustness of grasping and manipulation. But functional grasping with multi-fingered hands remains underexplored.

We propose learning functional grasps using low-quality reconstructed HOI data, without relying on costly human demonstrations. Specifically, we obtain hand-object interactions (HOI) from web-crawled images of humans holding objects using a pretrained 3D reconstruction model [21]. We retarget the human hand mesh to the multi-fingered ShadowHand robot using AnyTeleop [27], and align the noisy object meshes with accurate 3D shapes generated by text-to-3D tools Meshy AI [28]. Using HOI data, we train the interaction-centric grasp model DRO [4], which takes as input point clouds of the robot and object sampled from their respective meshes, and outputs a grasp joint configuration. To expand the web-based HOI dataset, we deployed the model trained on web data in IsaacGym and collected successful grasps through simulation. This simulator-augmented dataset further improves grasping performance. Overall, this paper makes the following contributions:

- We propose and demonstrate that reconstructed, low-quality hand-object interactions (HOI) from web images can be effectively leveraged to train dexterous and functional grasping policies.
- Our approach, trained on HOI data from web images and further augmented with simulation data, achieves grasp success rates of 61.8% and 83.4% respectively with the ShadowHand across a wide range of objects—including those underrepresented in existing datasets, such as syringes, pens, and spray bottles.
- Our approach outperforms baseline methods, achieving a 6.7% improvement in grasp success rate and a notable gain in human-evaluated functionality scores.

[*] denotes equal contribution.[1] Carnegie Mellon University, [2] Stanford University, [3] National University of Singapore.

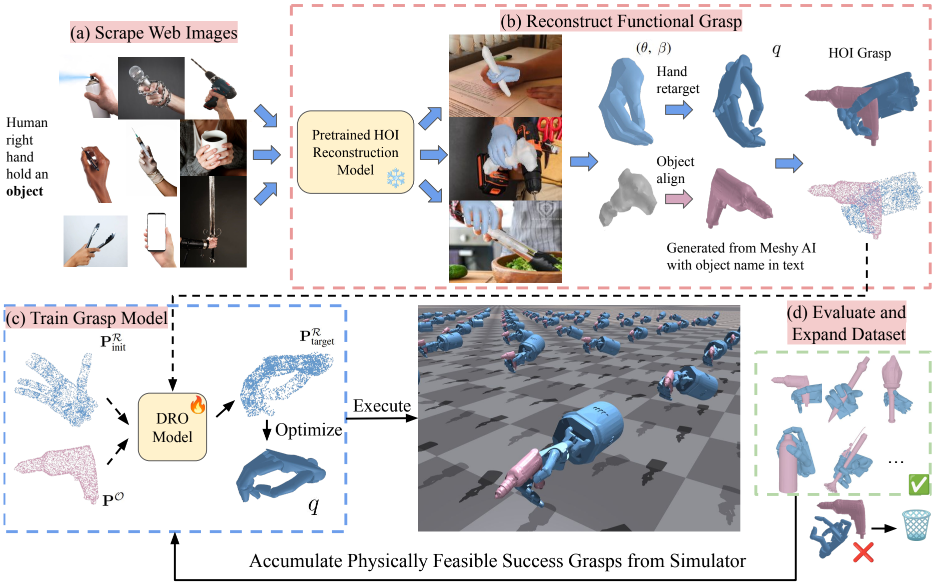
(a) Scrape Web Images
Human right hand hold an **object**
Pretrained HOI Reconstruction Model
(b) Reconstruct Functional Grasp
$(\theta,\ \beta)$
Hand retarget
q
HOI Grasp
Object align
Generated from Meshy AI with object name in text
(c) Train Grasp Model
$\mathbf{P}^{\mathcal{R}}_{\text{init}}$
DRO Model
$\mathbf{P}^{\mathcal{R}}_{\text{target}}$
Optimize
$\mathbf{P}^{\mathcal{O}}$
q
Execute
(d) Evaluate and Expand Dataset
...
Accumulate Physically Feasible Success Grasps from Simulator

Fig. 1: **Hand-Object Interaction (HOI) Collection, Grasp Model Training and Execution Pipeline.** (a) and (b): HOI grasp data, potentially containing penetrations and unrealistic contacts, is reconstructed from randomly scraped web images. (c): The DRO grasping model is trained on this HOI dataset to predict target joint configurations for grasp execution. (d): Physically feasible grasps are collected in simulation to expand the dataset and retrain the model.

II. Method

A. Reconstructing HOI as Grasp Training Data

Our method begins with the reconstruction of HOI from web-crawled images of humans holding objects, which are then used as a functional grasping training dataset. Given an image I depicting a hand holding an object, we aim to reconstruct the 3D shapes and poses of the hand and the object involved in the interaction. Specifically, we apply the method proposed by Ye et al. [21] to predict the MANO hand pose θ, which is a low-dimensional representation of human hand, using the off-the-shelf Frankmocap [29] and infer the shape of the interacting object in the same frame via a hand pose-conditioned signed-distance function (SDF).

To leverage reconstructed HOI data for dexterous grasp training, we first retarget the MANO hand poses to the robot joint configuration q. We employ AnyTeleop, which uses position-based optimization [27] to minimize the 3D positional error between keypoints on the robot's links and their corresponding points on the MANO hand mesh. The reconstructed object meshes are often of low quality, exhibiting imprecise geometry and visually unrealistic appearances—see the gray power drill in Figure 1(b)—particularly in cases of significant hand occlusion or when the objects are outside the training distribution [21, 22]. Thus, we align it with an accurate 3D mesh from the same category using the ICP algorithm, following the approach in [22]. These accurate 3D meshes can be easily obtained from online mesh generators by inputting the object name as text, such as MeshyAI [28] and Genie [30]. Our objective is not to achieve a pixel-perfect reconstruction of the object in the image, but to retain the key interaction cues—enabling the robot hand to interact with the object in a manner that is both natural and functionally aligned with human behavior. Figure 1 (a) and (b) shows the process.

B. Interaction-Centric Grasping Model Training

Although reconstructed HOI data is not as precise as human-collected data—often exhibiting issues such as excessive penetration or unstable finger contact—it can still capture the essential functionality of hand-object interaction, provided the reconstruction is accurate. As such, traditional robot-centric grasping models, which require accurate target joint configurations as training labels, or object-centric models, which rely on precise contact maps of the object shape, may not be ideal for this scenario. A recent alternative, the interaction-centric model DRO [4] measures relative distances between the robot hand and object point clouds and emphasize on their interaction, making it well-suited for imperfect HOI data.

In DRO, for a dexterous robot hand, such as the ShadowHand, and its URDF, we first sample points on the surface of each link, storing the resulting point clouds as $\{\mathbf{P}_{\ell_i}\}_{i=1}^{N_\ell}$, where N_ℓ is the number of links. Next, we define a point cloud forward kinematics model for the robot hand, $\text{FK}\left(q, \{\mathbf{P}_{\ell_i}\}_{i=1}^{N_\ell}\right)$, which maps each joint configuration to a corresponding set of point clouds. For instance, for an initial configuration q_{init}, we obtain $\mathbf{P}^{\mathcal{R}}_{\text{init}} = \text{FK}\left(q_{\text{init}}, \{\mathbf{P}_{\ell_i}\}_{i=1}^{N_\ell}\right) \in \mathbb{R}^{N_{\mathcal{R}} \times 3}$, where $N_{\mathcal{R}}$ is the total number of robot point clouds. Given $\mathbf{P}^{\mathcal{R}}_{\text{init}}$ and the object point cloud $\mathbf{P}^{\mathcal{O}} \in \mathbb{R}^{N_{\mathcal{O}} \times 3}$, sampled from the object mesh in the HOI data (where $N_{\mathcal{O}}$ represents the number of object point clouds), the objective of the DRO model is to predict the point-to-point distance matrix $\mathcal{D}(\mathcal{R}, \mathcal{O})^{\text{Pred}} \in \mathbb{R}^{N_{\mathcal{R}} \times N_{\mathcal{O}}}$. The training loss is computed by evaluating the difference between the predicted and ground-truth distance matrices with $\mathcal{L}_{\text{L1}}\left(\mathcal{D}(\mathcal{R}, \mathcal{O})^{\text{Pred}}, \mathcal{D}(\mathcal{R}, \mathcal{O})^{\text{GT}}\right)$.

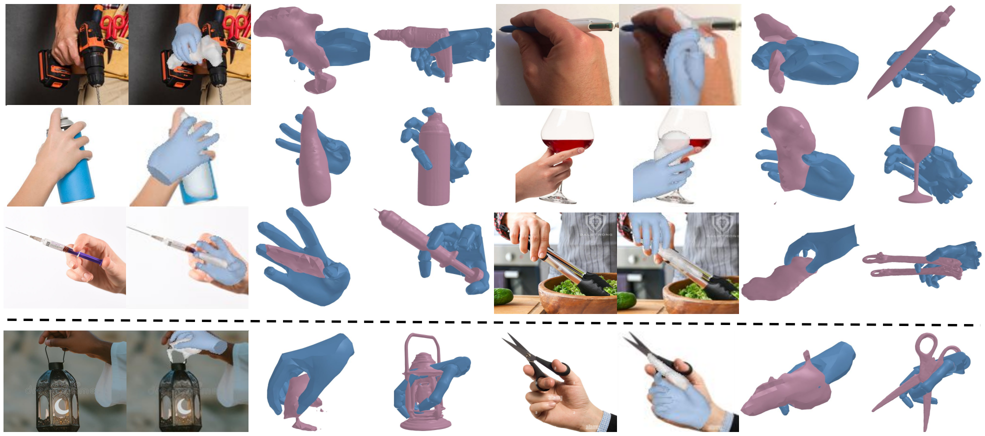

Fig. 2: **Visualization of HOI Reconstruction from Web Images.** Reconstructions for representative objects. **Success cases**—Row 1: *Power Drill*, *Pen*; Row 2: *Spray Bottle*, *Wine Glass*; Row 3: *Syringe*, *Tongs*. **Failure cases**—Row 4: *Lantern*, *Scissors*. From left to right: original web image, web image overlaid with HOI mesh, HOI mesh alone, and HOI mesh with retargeted robot hand and aligned object.

Using $\mathcal{D}(\mathcal{R},\mathcal{O})^{\text{Pred}}$ and the object point clouds $\mathbf{P}^{\mathcal{O}}$, we can position the robot point cloud in the target grasp pose $\mathbf{P}^{\mathcal{R}}_{\text{target}}$ with multilateration method [31]. Once the predicted grasp point cloud $\left\{\mathbf{P}^{\text{target}}_{\ell_i}\right\}_{i=1}^{N_\ell} \in \mathbf{P}^{\mathcal{R}}_{\text{target}}$ is obtained, we can compute the 6D pose of each link and the joint configuration through optimization. See Figure 1 (c) for visualization and see Wei et al. [4] for further details.

C. Simulator-Augmented Dataset

Although we can reconstruct thousands of HOI data from the vast collection of internet images, many of them are unusable due to issues such as improper hand detection, inaccurate object shape reconstruction, and faulty interactions. A significant portion must be filtered out—either automatically using numerical metrics like the object contact ratio [14], or manually through human inspection. Additionally, penetration between the hand and object, as well as unstable contact, presents further challenges. These issues can cause the DRO model trained on such data to apply excessive force, resulting in the object being bumped out during grasping, or fail to maintain a stable hold under external disturbances.

To improve both the quantity and quality of the grasping dataset, we collect functional, physically feasible training data through simulation evaluation. After the initial training phase with HOI data reconstructed from web images, we deploy the grasping model in IsaacGym [32], a high-fidelity physics simulator, to perform grasping tasks. Only successful grasps withstand external force distance are retained and added to the training set, shown in Figure 1 (d). This process ensures that the training grasps are functional and physically valid: the model learns from functional HOI examples, while the simulator filters out unstable or penetrative interactions.

III. Experiments

A. Simulation Grasp Experiment

1) Setup and Baselines.: We evaluate the grasp success rate using ShadowHand in IsaacGym simulator. Our method is compared against the following baselines: (1) *GenDexGrasp* [2]: Generates grasp poses for novel objects by leveraging contact maps as a hand-agnostic intermediate representation; (2) *DexGraspNet* [1]: A model trained on a large-scale dexterous grasp dataset; (3) *DRO* [4]: An interaction-centric model that predicts the relative distance between robotic hand and object point clouds. Our method adopts the DRO model architecture and trains it using our reconstructed HOI dataset (web data), which consists of 10 object categories with 100 web images per category: *Power Drill*, *Pen*, *Microphone*, *Phone*, *Spray Bottle*, *Wine Glass*, *Tong*, *Syringe*, *Mug*, *Sword*. We test on 9 unseen object categories: *Whip*, *Teapot*, *Axe*, *Remote*, *Torch*, *Hammer*, *Whisk*, *Hand Soap Bottle*, *Writing Brush*. For data augmentation, we deploy the model trained on web data in IsaacGym and collect 200 successful grasps per object—including those from the unseen test set—forming an expanded simulation dataset (sim data) for model retraining. For both the web and simulation datasets, models are trained on all object categories within their respective sets and evaluated based on grasp success rate. We define a successful grasp as one in which the object displacement remains below 2 mm under external force disturbances, indicating a stable grasp.

2) HOI Reconstruction Results.: (2) From a qualitative perspective, while not all reconstructed HOI data is accurate, many examples preserve the functional grasp patterns observed in the original web images—see the first three rows of Figure 2. The MANO hand pose detection tends to be relatively stable, but the reconstructed object meshes are often incomplete or capture only a coarse approximation of the object's shape (see the second and third columns of Figure 2). Although the aligned object mesh provides a more accurate geometry, several issues remain: (1) Penetration between the robot hand and the object exists, as observed in examples like *Spray Bottle* and *Syringe*. (2) Some contact points are imprecise—for example, the index finger is positioned between the prongs of the *Tongs*. Despite these limitations, we show in Section III-A.3 that a grasping model trained on this relatively low-quality data still performs effectively.

In addition, reconstructed hand-object interactions often fail for certain object types with complex geometry or interaction mode, such as buckets or scissors. For example, in the case of buckets, the reconstruction typically captures the

Dataset	Power drill	Pen	Microphone	Phone	Spray bottle	Wine glass	Tong	Syringe	Mug	Sword
Web data	98	92	98	62	74	86	72	64	88	24
Sim aug	92	97	99	80	71	92	94	72	99	55
	Whip	Teapot	Axe	Remote	Torch	Hammer	Whisk	Soap bottle	Writing brush	Average
Web data	8	12	10	92	82	94	20	80	19	61.8
Sim aug	30	99	82	89	96	78	100	82	78	**83.4**

TABLE I: **Grasp Success Rates in IsaacGym for Models Trained on Different Datasets.** Objects from *Power Drill* to *Sword* are seen during training, while *Whip* to *Writing Brush* are unseen. Both methods are evaluated over 100 trials per object. The model trained on web data performs well on seen objects, and performance further improves on unseen objects with simulation-augmented data.

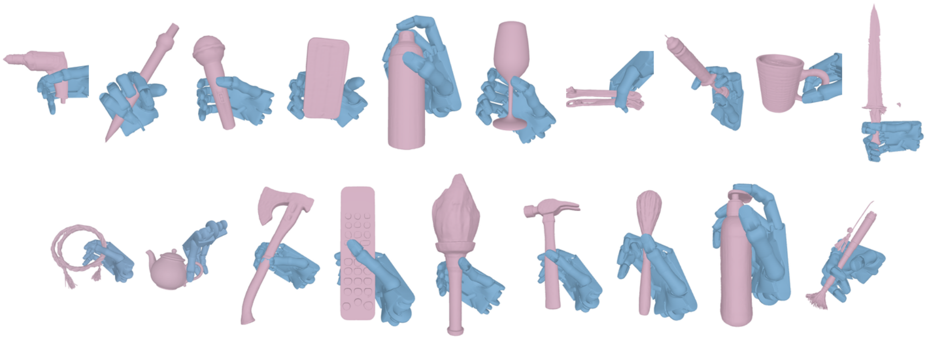

Fig. 3: **Visualization of Generated Grasps Across All Objects.** The first row shows objects seen during web-data training, while the second row presents unseen objects from the test set.

main body while missing thin structures like the handle. Consequently, after ICP alignment, the hand is positioned near the bucket's body instead of grasping the handle. Scissors also present challenges due to their unique interaction mode, where fingers are placed inside the holes of handles—a pattern not commonly found in other objects. See the last row of Figure 2 for an illustration.

3) Simulation Grasp Results.: (1) Our grasping model, trained with web data and simulation-augmented datasets, achieves success rates of 61.8 % and 83.4 %, respectively, as summarized in Table I. This demonstrates that functional grasping models can be effectively trained using 3D reconstructed HOI from web data, without requiring extensive human annotations. Although performance on certain seen objects (e.g., *Phone*, *Long Sword*) and unseen objects may initially be lower due to limited data, it improves significantly with the addition of simulated interaction data generated in IsaacGym using the web-trained model. For instance, we observe frequent hand-object penetrations when grasping the *Phone*, and difficulty in securely holding the *Sword* in web dataset, both of which negatively affect success rates. The simulation-augmented dataset helps address these issues by collecting physically plausible grasps since IsaacGym prevents penetration and ensures that only stable contacts, capable of withstanding external force disturbances, are saved. Moreover, the model trained with simulation data demonstrates functional grasp behavior, see Figure 3.

	SR (%)	Functionality
GenDexGrasp [2]	20.56	6.4
DexGraspNet [1]	39.56	19.6
DRO [4]	33.44	26.3
Our (Web data)	**46.33**	**47.7**

TABLE II: **Grasping Performance of Unseen Objects in Simulation.** SR = Success Rate. Functionality scores are based on human evaluation of functional correctness.

(2) Our grasping model outperforms all baselines in both success rate and functionality score, as summarized in Table II. Functionality is evaluated through human studies, where participants are shown two grasps—each from a randomly selected pair of methods—and asked to choose the one they find more natural and functional. We normalize the number of votes received by each method to compute the final scores. For a fair comparison, all methods are evaluated on the same set of 9 unseen test objects. While baseline methods perform well on common objects that require only stable grasps—as shown in their original papers—they often overlook functional grasp strategies, such as placing a finger on top of a *Spray Bottle* for operate. In contrast, our method learns such nuanced interactions from web data and generalizes effectively to novel but related object categories, such as transferring from *Pen* to *Writing Brush*, *Mug* to *Teapot*, and *Spray Bottle* to *Hand Soap Bottle*.

IV. CONCLUSIONS

This work introduces and demonstrates an approach for learning functional grasps using 3D reconstructed HOI data from internet images, bypassing the need for expensive human-collected data. The method integrates an interaction-centric grasping model DRO to focus on the functional interactions between hand and object point clouds, making it effective for noisy HOI data. Experiments show that the proposed approach can: (i) effectively handle low-quality HOI data to train high-performance functional grasping models, (ii) be tested across a wide range of objects requiring diverse functional grasps, using the Shadowhand in simulation, and (iii) outperform baseline models in both grasping success rates and human-evaluated functionality.

References

[1] R. Wang et al. “Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation”. In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2023, pp. 11359–11366.

[2] P. Li et al. “Gendexgrasp: Generalizable dexterous grasping”. In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2023, pp. 8068–8074.

[3] T. Liu et al. “Synthesizing diverse and physically stable grasps with arbitrary hand structures using differentiable force closure estimator”. In: *IEEE Robotics and Automation Letters* 7.1 (2021), pp. 470–477.

[4] Z. Wei et al. “D (R, O) Grasp: A Unified Representation of Robot and Object Interaction for Cross-Embodiment Dexterous Grasping”. In: *arXiv preprint arXiv:2410.01702* (2024).

[5] H. Zhang et al. “RobustDexGrasp: Robust Dexterous Grasping of General Objects from Single-view Perception”. In: *arXiv preprint arXiv:2504.05287* (2025).

[6] W. Wan et al. “Unidexgrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 3891–3902.

[7] Y. Qin et al. “Dexpoint: Generalizable point cloud reinforcement learning for sim-to-real dexterous manipulation”. In: *Conference on Robot Learning*. PMLR. 2023, pp. 594–605.

[8] S. Li et al. “Shapegrasp: Zero-shot task-oriented grasping with large language models through geometric decomposition”. In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2024, pp. 10527–10534.

[9] K. Shaw, S. Bahl, and D. Pathak. “Videodex: Learning dexterity from internet videos”. In: *Conference on Robot Learning*. PMLR. 2023, pp. 654–665.

[10] S. Brahmbhatt et al. “Contactgrasp: Functional multi-finger grasp synthesis from contact. In 2019 IEEE”. In: *RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2019, pp. 2386–2393.

[11] A. Agarwal et al. “Dexterous functional grasping”. In: *arXiv preprint arXiv:2312.02975* (2023).

[12] H. G. Singh et al. “Hand-object interaction pretraining from videos”. In: *arXiv preprint arXiv:2409.08273* (2024).

[13] S. Brahmbhatt et al. “Contactdb: Analyzing and predicting grasp contact via thermal imaging”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 8709–8719.

[14] L. Yang et al. “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, pp. 20953–20962.

[15] O. Taheri et al. “GRAB: A dataset of whole-body human grasping of objects”. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer. 2020, pp. 581–600.

[16] X. Zhan et al. “Oakink2: A dataset of bimanual hands-object manipulation in complex task completion”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 445–456.

[17] W. Wei et al. “Learning human-like functional grasping for multi-finger hands from few demonstrations”. In: *IEEE Transactions on Robotics* (2024).

[18] M. Aburub et al. “Functional Eigen-Grasping Using Approach Heatmaps”. In: *IEEE Robotics and Automation Letters* (2025).

[19] G. Pavlakos et al. “Reconstructing hands in 3d with transformers”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 9826–9836.

[20] B. Wen et al. “Foundationpose: Unified 6d pose estimation and tracking of novel objects”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 17868–17879.

[21] Y. Ye, A. Gupta, and S. Tulsiani. “What’s in your hands? 3d reconstruction of generic objects in hands”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 3895–3905.

[22] J. Wu et al. “Reconstructing Hand-Held Objects in 3D from Images and Videos”. In: *arXiv preprint arXiv:2404.06507* (2024).

[23] Y. Liu et al. “EasyHOI: Unleashing the Power of Large Models for Reconstructing Hand-Object Interactions in the Wild”. In: *arXiv preprint arXiv:2411.14280* (2024).

[24] Y. Ye et al. “Affordance diffusion: Synthesizing hand-object interactions”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 22479–22489.

[25] L. Wang et al. “A real2sim2real method for robust object grasping with neural surface reconstruction”. In: *2023 IEEE 19th International Conference on Automation Science and Engineering (CASE)*. IEEE. 2023, pp. 1–8.

[26] M. Torne et al. “Reconciling reality through simulation: A real-to-sim-to-real approach for robust manipulation”. In: *arXiv preprint arXiv:2403.03949* (2024).

[27] Y. Qin et al. “AnyTeleop: A General Vision-Based Dexterous Robot Arm-Hand Teleoperation System”. In: *Robotics: Science and Systems*. 2023.

[28] Meshy AI. *Meshy AI: The #1 AI 3D Model Generator for Creators*. Accessed: 2025-04-17. 2025. URL: `https://www.meshy.ai/`.

[29] Y. Rong, T. Shiratori, and H. Joo. “Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration”. In: *arXiv preprint arXiv:2008.08324* (2020).

[30] L. AI. *Genie: Text-to-3D AI Model*. Accessed: 2025-04-17. 2025. URL: `https://lumalabs.ai/genie?view=create`.

[31] A. Norrdine. “An algebraic solution to the multilateration problem”. In: *Proceedings of the 15th international conference on indoor positioning and indoor navigation, Sydney, Australia*. Vol. 1315. 2012.

[32] V. Makoviychuk et al. “Isaac gym: High performance gpu-based physics simulation for robot learning”. In: *arXiv preprint arXiv:2108.10470* (2021).