Guiding Unified Dexterous Grasp Synthesis Across Modes and Scales via Learned Human Priors

Anonymous Author(s)
Affiliation
Address
email

Figure 1: **Human priors guide scalable dexterous grasp synthesis across modes and scales.** HUGS learns an object-conditioned human prior from a compact self-collected dataset to predict preferred contact modes and wrist initializations. Guided by this prior, force-closure-aware optimization synthesizes diverse and stable grasps ranging from two-finger pinches to bimanual grasps.

Abstract: Dexterous grasping across diverse object scales requires contact modes ranging from two-finger pinches to bimanual grasps. Existing dexterous grasp synthesis methods reduce the high-dimensional optimization space with manually designed expected contacts and initialization heuristics, which struggle to balance synthesis success rate and diversity. We present **HUGS**, a Human-prior-guided framework for Unified dexterous Grasp Synthesis across modes and scales. Instead of directly retargeting human demonstrations, HUGS learns an object-conditioned human prior that captures human grasp preferences and guides downstream force-closure-aware optimization. The prior is trained on a compact self-collected human grasp dataset with 1.8K grasps over 304 objects, providing broad coverage of object scales and contact modes. During synthesis, HUGS adaptively proposes contact modes and wrist initializations, substantially improving the balance between contact-mode coverage and synthesis success rate over heuristic-based methods. With HUGS, we synthesize 3.2M robotic grasps over 157K scenes, spanning object half-diagonal lengths from 2 cm to 30 cm and modes from two-finger to bimanual

Submitted to the 10th Conference on Robot Learning (CoRL 2026). Do not distribute.

grasps. Models trained on the synthesized dataset autonomously select appropriate contact modes in the real world, enabling grasping from screws to large boxes.

Keywords: Dexterous Grasping, Human Priors, Grasp Synthesis

1 Introduction

Dexterous grasping is inherently multi-mode across object scales, from two-finger precision grasps to coordinated bimanual grasps, and the same object may admit different modes under different conditions. Training a generalizable grasp generation model for such diversity requires massive datasets, yet large-scale real-world collection is impractical, making scalable grasp synthesis essential.

Dexterous grasp synthesis involves a vast search space due to the high degrees of freedom and multi-contact nature. Existing methods often constrain optimization with manually predefined contact modes, such as single-hand [1, 2] or dual-hand full-finger grasps [3, 4], and heuristic wrist pose initializations. However, these coarse rules fail to exploit object-specific information and struggle to balance synthesis efficiency and grasp diversity: loose heuristics lead to low-quality optimization and inefficient synthesis, while restrictive heuristics severely limit grasp diversity.

Indeed, anthropomorphic robotic hands are designed for human-like manipulation [5], making human grasps a natural source of knowledge for robots. Existing methods often retarget human grasps to robotic hands [6, 7, 8], converting each demonstration into a robot grasp for the same or highly similar object and thereby limiting scalable synthesis over large-scale, diverse object sets.

Our key insight is that we can use human demonstrations to learn generalized and reusable **object-conditioned human priors**, instead of directly retargeting each human grasp to a robotic hand in a one-to-one manner. By capturing human preferences for specific object geometry, the prior guides robotic grasp optimization with better global initializations and optimization targets for efficient convergence to higher-quality solutions. Such a **Human Prior + Robot Optimization** paradigm offers following advantages: **1) Adaptivity**: compared with hand-crafted heuristics, the learned human prior provides object-aware global guidance that constrains optimization to more suitable regions of search space. **2) Scalability**: unlike direct retargeting methods, the learned prior generalizes to unseen objects, enabling scalable synthesis of large and diverse robotic grasp datasets from limited human demonstrations. and **3) Physical Plausibility**: force-aware optimization accounts for robot-specific kinematics and task physics, while tolerating slight errors in coarse human prior.

In this work, we present **HUGS**, a Human-prior-guided framework for Unified dexterous Grasp Synthesis across modes and scales. To learn a generalizable human prior from limited data, we abstract human preference into contact modes and wrist poses, the two components most critical for robotic grasp synthesis. Using a self-collected dataset with 304 objects and 1.8K grasps, we train an object-conditioned generative model that predicts human-preferred grasp configurations. During synthesis, the learned prior proposes object-adaptive contact modes and wrist initializations, enabling scalable and efficient synthesis of 3.2M grasps over 157K scenes, spanning objects with half-diagonal lengths from 2 cm to 30 cm and contact modes from two-finger to bimanual grasps, forming what is, to our knowledge, the first large-scale dataset with multiple modes per object (Fig. 1). Models trained on the dataset adaptively select suitable contact modes, enabling grasping from screws to large boxes in the real world. Our contributions are highlighted as follows:

1. **Human Grasp Dataset Across Modes and Scales**. We collect and publicly release a compact yet diverse human grasp dataset spanning a broad range of object scales and contact modes, providing a strong foundation for learning generalizable human grasp priors.
2. **Human-Prior-Guided Unified Grasp Synthesis**. We propose a human-prior-guided framework that learns object-conditioned priors over contact modes and wrist poses for unified dexterous grasp synthesis across scales and modes, eliminating manually designed coarse heuristics.
3. **Large-Scale Synthesis and Evaluation**. We demonstrate scalable synthesis of diverse dexterous grasp datasets and systematically evaluate the framework in terms of cross-scale and cross-mode prior prediction, synthesis quality, efficiency, diversity, and real-world grasping potential.

2 Related Work

Dexterous Grasp Synthesis. Analytical grasp synthesis methods optimize hand configurations with physics-based objectives such as force closure. To reduce the search space of high-DoF multi-contact systems, existing methods typically rely on manually predefined contact regions and heuristic wrist initialization strategies. Most prior work focuses on single-hand full-finger grasps [9, 1, 2, 10] or bimanual grasps [3, 4]. Different contact modes can be synthesized by changing predefined contact regions, e.g., two-finger [11, 12] or three-finger grasps [13, 12]. However, these modes are typically manually tied to object-scale ranges, causing each object scale to correspond to only one grasp mode and limiting contact-mode diversity. In addition, most methods randomly sample wrist poses around the object's surface [1, 2, 3]. For bimanual grasping, random sampling ignores inter-hand coordination, while strict symmetry constraints restrict diversity [3, 12]. Hand-centric methods [14, 15] improve diversity and efficiency for single-hand floating grasps, but may be less suitable for bimanual grasping and environment-constrained settings such as tabletop grasping.

Human Data for Grasp Synthesis. Human demonstrations are widely used for functional grasp synthesis through retargeting-based approaches that transfer human grasps to robotic hands. Some methods directly retarget each human grasp to a robotic grasp [7, 16], while others synthesize grasps for similar objects from a small set of human grasp templates [6, 17, 18, 8]. However, limited by the scale of human functional grasp datasets, these approaches typically support only limited object categories and rely on similarity to existing templates, rather than generalizing to arbitrary objects. Some works learn object-conditioned contact maps from human data [19, 20], but remain limited to single-hand full-finger grasps and require highly accurate human contact predictions that may not align with robotic hand kinematics. In contrast, our method uses only high-level human priors, namely contact modes and wrist poses, while retaining sufficient flexibility for force-closure grasp optimization compatible with robotic hand kinematics.

Online Grasp Generation. Synthetic grasp datasets are widely used to train grasp generation networks from partial observations such as images or point clouds for real-time deployment [21, 22, 23, 13], including grasp generation in cluttered environments [24, 25, 26]. A few works directly learn grasping policies with reinforcement learning, bypassing explicit grasp pose planning [27, 28, 29, 30, 31]; however, they often suffer from limited grasp diversity and unnatural finger motions.

3 Method

3.1 Problem Formulation

Grasp Optimization Problem. Let o denote the target object with surface $\mathcal{S}_o$. For hand $h \in \{L, R\}$, let $\boldsymbol{q}^h \in \mathbb{R}^{d_q}$ be the joint configuration and $\boldsymbol{T}^h = (\boldsymbol{R}^h, \boldsymbol{t}^h) \in SE(3)$ be the wrist pose, with bimanual states $\boldsymbol{Q} = (\boldsymbol{q}^L, \boldsymbol{q}^R)$ and $\boldsymbol{T} = (\boldsymbol{T}^L, \boldsymbol{T}^R)$. A grasp $g = (\boldsymbol{Q}, \boldsymbol{T}, c)$ uses contact configuration c to specify which hand regions contact the object, and grasp synthesis models $\pi(g \mid o) = \pi(\boldsymbol{Q}, \boldsymbol{T}, c \mid o)$. Active contact regions c induce contact points $\{\boldsymbol{p}_i\}_{i=1}^{N_c}$ with normals $\{\boldsymbol{n}_i\}_{i=1}^{N_c}$. Under point contact with Coulomb friction, each contact force satisfies $\boldsymbol{f}_i \in \mathcal{F}_i = \{\boldsymbol{f}_i \mid \|\boldsymbol{f}_i^t\| \le \mu f_i^n,\ f_i^n \ge 0\}$ and induces wrench $\boldsymbol{w}_i = [\boldsymbol{f}_i;\ (\boldsymbol{p}_i - \boldsymbol{x}_{\text{com}}) \times \boldsymbol{f}_i]$, where $\boldsymbol{x}_{\text{com}}$ is the object center of mass. The grasp wrench space is $\mathcal{G}(g) = \{\sum_{i=1}^{N_c} \boldsymbol{w}_i \mid \boldsymbol{f}_i \in \mathcal{F}_i\}$, and force closure holds when $0 \in \text{int}(\mathcal{G}(g))$. We formulate grasp optimization as minimizing force-closure energy $\Phi(g)$ under feasibility constraints:

$$\min_{g=(\boldsymbol{Q},\boldsymbol{T},c)} \Phi(g) \quad \text{s.t.} \quad \boldsymbol{Q} \in [\boldsymbol{Q}_{\min}, \boldsymbol{Q}_{\max}],\ \boldsymbol{p}_i \in \mathcal{S}_o\ \forall i,\ \text{CollisionFree}(g, o). \tag{1}$$

Here $\boldsymbol{p}_i$ are induced by active regions in c; the constraints enforce joint limits, object-surface contacts, and collision-free hand-object and hand-hand configurations.

Overview of HUGS. In grasp pose optimization, the contact configuration c determines the number and locations of contact points, making the problem hybrid discrete-continuous. Existing methods typically reduce the combinatorial search space with predefined contact regions. Meanwhile, the wrist pose $\boldsymbol{T}$ is highly global, and poor initializations often trap local optimization in suboptimal minima.

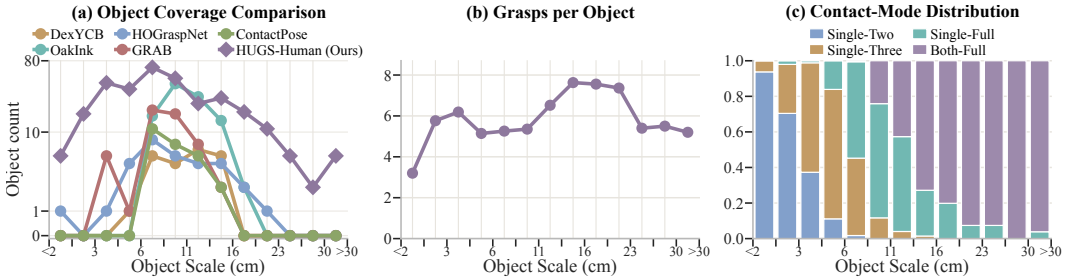


Figure 2: **Statistics of HUGS-Human Dataset.** (a) Object-count distribution, compared with existing human grasp datasets (OakInk: only real-world captured objects; log-scaled y-axis and power-scaled x-axis, with absolute-value labels). (b) Average number of HUGS-Human grasps per object. (c) Contact-mode distribution in HUGS-Human.

Consequently, optimization quality largely depends on initializing c and $\boldsymbol{T}$, while optimizing the hand joint configuration $\boldsymbol{Q}$ is relatively straightforward given suitable c and $\boldsymbol{T}$. Existing methods mainly use fixed coarse heuristics to initialize contact configurations and wrist poses, wasting optimization budget on implausible grasp modes while missing object-aware strategies. HUGS instead learns an object-conditioned human grasp prior to organize the search space before robot-specific optimization.

We adopt a compact discrete contact-mode representation with four coarse modes, selected as prevalent human grasping strategies that current dexterous robot hands can execute: 1) **Single-Two**, a single-hand two-finger grasp for small objects with limited contact area; 2) **Single-Three**, a single-hand three-finger grasp, the minimum for force closure under frictional point contacts [32]; 3) **Single-Full**, a denser single-hand full-finger grasp requiring larger accessible object surface area; and 4) **Both-Full**, a full-finger bimanual grasp for large objects requiring broader surface coverage. Based on this contact-mode abstraction, we formulate the grasp synthesis as

$$\pi(\boldsymbol{Q}, \boldsymbol{T}, c \mid o) = \pi(\boldsymbol{Q}, \boldsymbol{T} \mid \boldsymbol{T}_0, c, o)\, \pi(\boldsymbol{T}_0 \mid c, o)\, \pi(c \mid o), \tag{2}$$

where the learned prior predicts $\pi(c \mid o)$ and $\pi(\boldsymbol{T}_0 \mid c, o)$ to propose plausible contact modes c and wrist initializations $\boldsymbol{T}_0$. Conditioned on these high-level proposals, the robot optimizer solves for $\boldsymbol{Q}$ and refines $\boldsymbol{T}$ under the force-closure and feasibility constraints in Eq. 1.

3.2 Compact Human Grasp Dataset

Dataset Motivation. Existing human grasp datasets mainly target hand-object reconstruction rather than synthesis guidance. Most emphasize single-hand grasps with limited scale coverage, such as DexYCB [33], OakInk [34], and HOGraspNet [35]; those containing bimanual grasps, such as GRAB [36] and ContactPose [37], include them only sparsely. They also do not explicitly consider mode and wrist-pose diversity, making them less suitable for learning priors over c and $\boldsymbol{T}_0$ across scales.

Self-Collected Dataset. We introduce **HUGS-Human**, a compact human grasp dataset designed to learn the high-level prior used by HUGS. Participants grasp and lift tabletop objects from multiple supporting poses. For each posed object, we encourage diverse human-plausible contact modes and wrist poses while avoiding duplicates. The dataset contains 304 canonical objects, 525 posed objects, and 1.8k distinct human grasps. The collected grasps reveal three consistent patterns: 1) humans adapt contact modes and wrist poses according to object geometry and scale; 2) a single object can support multiple plausible contact modes, especially when its geometry is asymmetric; and 3) contact modes and feasible wrist poses are strongly coupled, as certain wrist poses only permit sparse fingertip contacts while others enable stable full-hand grasps on the same object. These patterns motivate our object-conditioned prior over c and $\boldsymbol{T}_0$. Each collected data sample includes the object mesh, object pose, hand wrist pose, MANO pose [38], and contact mode. The setup and annotation details are provided in the appendix. As shown in Fig. 2, HUGS-Human provides broad object-scale coverage compared with existing datasets, especially for very small and large objects, maintains multiple grasps per object across object scale bins, and captures the scale-dependent transition from sparse single-hand modes to full-hand and bimanual modes.

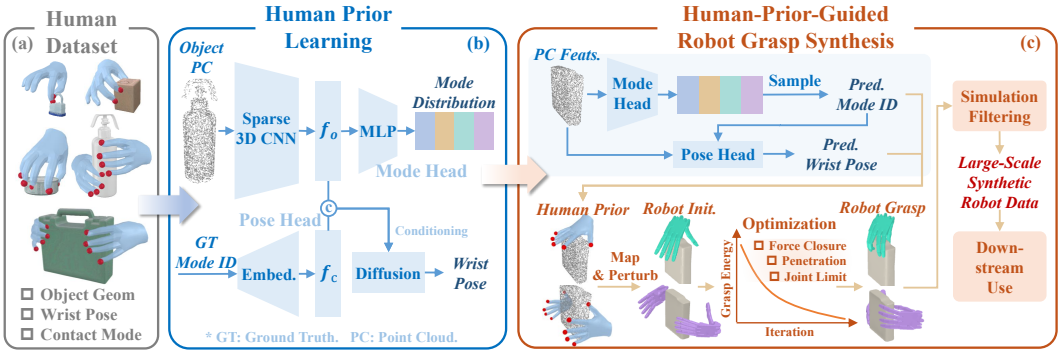


Figure 3: **Overview of HUGS.** HUGS learns an object-conditioned human prior from demonstrations with geometry, wrist poses, and contact modes. During synthesis, it predicts contact-mode distributions and wrist initializations for unseen objects, transfers them to robot hands, and optimizes force closure under feasibility constraints to synthesize large-scale robot grasps across modes and scales.

3.3 Object-Conditioned Human Grasp Prior

Prior Formulation. We learn object-conditioned distributions over contact modes and wrist poses, denoted by $\pi(c \mid o)$ and $\pi(\boldsymbol{T}_0 \mid c, o)$. Because human and robot wrist frames may differ, $\boldsymbol{T}_0$ uses the index metacarpophalangeal (MCP) position for translation and the dorsal-hand wrist orientation for rotation. We interpret these distributions as a **synthesis-budget prior**: for each object, they specify how much optimization effort to allocate to each contact mode and wrist-pose region. Implausible modes receive near-zero budget, while plausible modes share the budget according to human grasp diversity. This guides synthesis toward diverse plausible grasps with fewer optimization attempts.

Architecture and Training. As shown in Fig. 3(b), the prior takes a centered tabletop-frame object point cloud and encodes it into $\boldsymbol{f}_o$ using a MinkowskiEngine Sparse3DConv backbone [39]. A mode head predicts $\pi(c \mid o)$ with an MLP and softmax, supervised by the empirical contact-mode distribution of demonstrations for the same posed object using soft-label cross entropy. A mode-conditioned diffusion head models $\pi(\boldsymbol{T}_0 \mid c, o)$ by conditioning on $\boldsymbol{f}_v = [\boldsymbol{f}_o, \boldsymbol{f}_c]$, where $\boldsymbol{f}_c$ is a learnable mode embedding, and is trained on individual data samples with contact-mode labels and human wrist poses. We use random vertical rotations and point-cloud noise for augmentation.

3.4 Human-Prior-Guided Grasp Synthesis

Synthesis Overview. During grasp synthesis, the learned human prior predicts $\pi(c \mid o)$ and $\pi(\boldsymbol{T}_0 \mid c, o)$. To account for the size differences between human and robot hands, we query the prior with an object point cloud rescaled by the human-to-robot hand scale ratio. Given sampled modes and wrist initializations, optimization produces robot grasps following $\pi(\boldsymbol{Q}, \boldsymbol{T} \mid \boldsymbol{T}_0, c, o)$. Together, they induce the final grasp distribution $\pi(\boldsymbol{Q}, \boldsymbol{T}, c \mid o)$ for cross-mode synthesis. Finally, we filter grasps in MuJoCo [40] and retain all validated grasps across contact modes for each object, rather than only the best one [12], to preserve dataset diversity for downstream applications.

Prior-Guided Optimization Budget Allocation. Before optimization, the learned prior predicts $\pi(c \mid o)$ and samples wrist initializations for each contact mode. For each mode, we draw $4K$ candidates, rank them by approximate diffusion likelihood, keep the top $2K$, and apply farthest-pose selection to retain K diverse human-like initializations. The retained poses are mapped to the robot hand and slightly perturbed before optimization. Given a total budget of B optimization attempts, we allocate the mode budget as $B_c = \pi(c \mid o)B$ with a maximum per-mode cap. Each attempt samples one retained wrist initialization from its mode and open-hand joint angles.

Grasp Optimization Formulation. Given a contact mode c and wrist initialization $\boldsymbol{T}_0$, we follow [2] and formulate grasp synthesis as a bi-level optimization, where the outer problem updates $(\boldsymbol{Q}, \boldsymbol{T})$ and the inner problem evaluates force closure. We instantiate the force-closure energy in Eq. 1 as $\Phi(g) = \sum_{j=1}^{s} \Phi_j(g)$ over disturbance directions $\{\boldsymbol{t}_j\}_{j=1}^{s}$, where each directional residual is the

optimal value of the following quadratic program (QP):

$$\Phi_j(g) \triangleq \min_{\boldsymbol{F}_j} \left\| \beta \boldsymbol{t}_j - \sum_{i=1}^{N_c} \boldsymbol{G}_i(g) \boldsymbol{f}_{j,i} \right\|^2 \quad \text{s.t. } \boldsymbol{f}_{j,i} \in \mathcal{F}_i,\ i = 1, \ldots, N_c;\ \sum_{i=1}^{N_c} f_{j,i}^n \geq \gamma. \tag{3}$$

Here, $\boldsymbol{F}_j = \{\boldsymbol{f}_{j,i}\}_{i=1}^{N_c}$ are contact forces and $\boldsymbol{G}_i(g)$ is grasp matrix of contact point $\boldsymbol{p}_i$; Φ_j measures resistance to disturbance $\boldsymbol{t}_j$. Unlike [12, 4], we use a weighted sum of global and per-hand force-closure terms for bimanual synthesis to promote both bimanual stability and individual hand quality.

4 Results

Evaluation Goals. We evaluate whether our object-geometry-aware human prior improves large-scale multi-mode synthesis through better contact mode allocation and wrist initialization, and supports downstream model distillation for real-world deployment across object scales.

Implementation and Baselines. We construct object scenes from DGN2k [2], which includes 2.4k meshes. For each object, we sample multiple tabletop-stable poses and rescale it to 12 target sizes with AABB half-diagonal lengths from 2 cm to 30 cm, yielding 157k scenes spanning diverse geometries, poses, and scales. Each scene uses $B = 40$ optimization attempts in total, with at most $B_c = 20$ attempts per mode. Main results are reported on the Shadow Hand, with LEAP Hand [41] results in the appendix. Averaged over contact modes and object scales, HUGS optimizes 28.6 grasps per second on an RTX 4090. We compare HUGS against heuristic and ablated variants that differ in how contact modes and wrist-pose initializations are allocated. For the heuristic variants, scalar-scale contact-mode rules are derived from our human grasp dataset, and wrist poses are randomly initialized by convex-hull sampling with palm-facing poses. Further details are provided in the appendix.

- **Heur-Fix.** Uses the fixed *Single-Full* contact mode for every target object and scale, similar to [2].
- **Heur-Single.** Assigns one hand-designed contact mode to each object scale bin, similar to [12].
- **Heur-Multi.** Assigns multiple hand-designed contact modes to each object scale bin.
- **HUGS-Single.** Uses the HUGS wrist-pose prior, but allows only one mode per scale bin.
- **HUGS.** Our full method uses the human prior to allocate modes and initialize wrists.

4.1 Does Human Prior Predict Contact Modes Better than Scalar-Scale Rules?

We evaluate whether the geometry-aware human prior predicts contact-mode preferences better than rules based only on scalar object scale. HUGS-Human objects are split 8:2 into train and test sets; scalar-scale baselines map the object diagonal length to mode distributions using training-set statistics. Table 1 reports KL divergence, soft precision, and soft recall on test set, with metric and baseline details in the appendix. The human prior achieves lower KL and higher precision and recall, suggesting that contact modes depend on object geometry beyond scalar scale.

Table 1. Comparison of contact-mode prediction on held-out objects.

Method	KL ↓	Prec. ↑	Rec. ↑
Scale Rules	0.612	0.805	0.892
Human Prior	**0.300**	**0.873**	**0.964**

4.2 Does Human-Prior Guidance Improve Efficient Multi-Mode Grasp Synthesis?

We next evaluate how human-prior guidance improves synthesis. Fig. 4(a) shows per-mode budgets (light background bars) and success counts (dark foreground bars), and Fig. 4(b) reports overall success across object scales. The heuristic baselines reveal the limits of scale-only rules. **Heur-Fix** works only near the scale suited to *Single-Full*, showing the need for contact-mode adaptation. **Heur-Single** is more robust, but one scale-dependent mode cannot cover multiple valid strategies within the same scale. **Heur-Multi** adds mode diversity, yet wastes attempts on object-specific mismatches, such as single-hand attempts on 13–23 cm objects. **HUGS** instead allocates budget by object geometry, smoothly shifting from *Single-Two* to *Single-Three*, *Single-Full*, and finally *Both-Full* as scale grows. It also captures object-specific exceptions within the same scale: at 13–19 cm, HUGS makes fewer *Single-Full* attempts but targets the few objects that admit single-hand grasps more accurately,

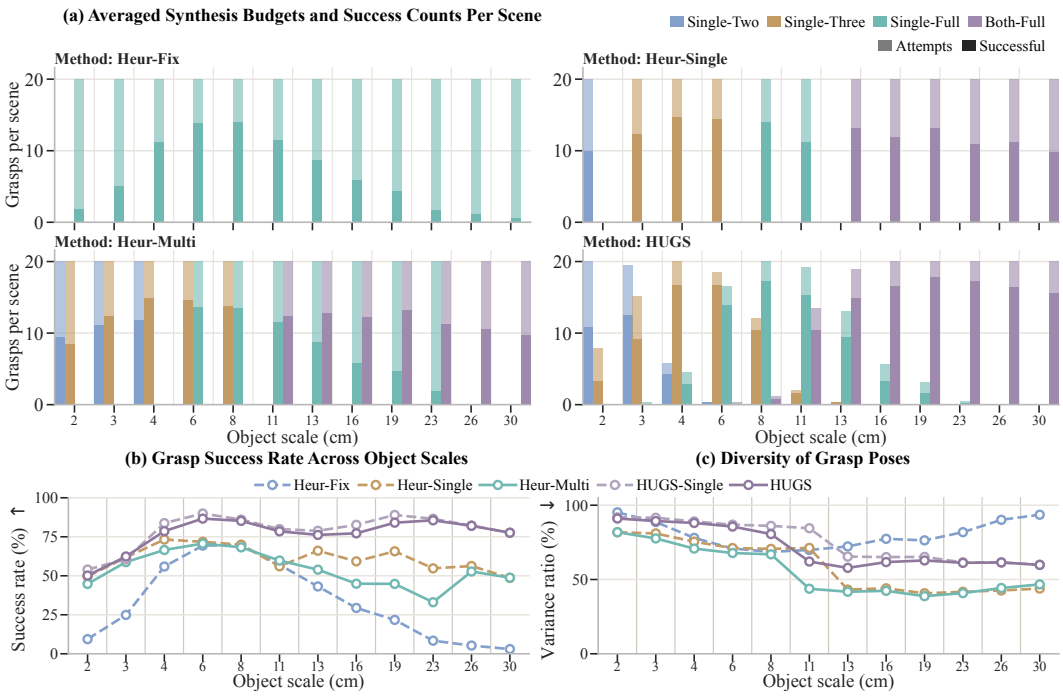


Figure 4: **Contact-mode allocation and synthesis success across object scales.** (a) Averaged synthesis budgets and success counts per scene. (b) Overall synthesis success rate across object scales. (c) Pose diversity measured by the explained-variance ratio of the first principal component.

yielding a much higher success rate than Heur-Multi. Thus, HUGS consistently outperforms all baselines. Fig. 1 visualizes diverse synthesized grasps, where many adjacent examples show the same object synthesized with different contact modes. Separately, HUGS-Single outperforms Heur-Single under the same single-mode setting, isolating the benefit of human-prior wrist initialization.

Fig. 4(c) compares grasp diversity using the explained-variance ratio of the first principal component [2, 14]. HUGS has lower pose dispersion because it concentrates on human-preferred wrist regions. This is expected: heuristic sampling increases diversity partly via unnatural reversed wrist poses (Fig. 5), which are difficult for humanoid robots to execute.

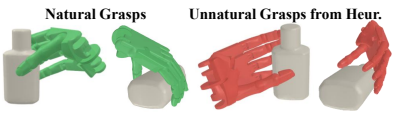


Figure 5: Heuristic wrist-pose sampling produces unnatural grasps.

4.3 Can Synthesized Grasps Supervise Online Grasp Generators?

We further test whether synthesized grasps can supervise learning-based grasp generation. We train identical lightweight object-conditioned generators on Heur-Multi and HUGS data, both of which contain multiple contact modes for the same object, using an 8:2 DGN2k object split and evaluating on held-out scenes across scales. The generator predicts binary contact-mode availability and corresponding grasp poses; architecture details are in the appendix. With HUGS data, contact-mode

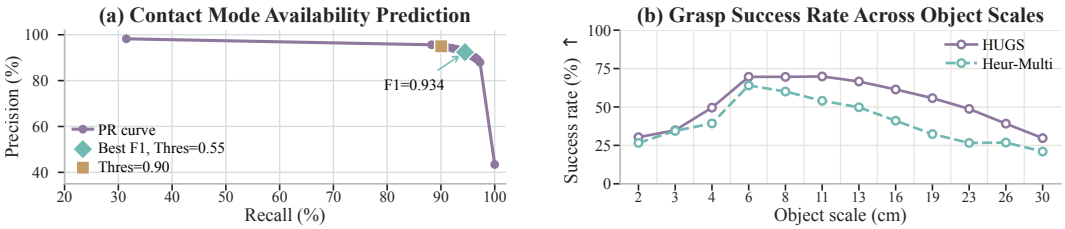


Figure 6: **Distilling from synthesized grasps.** (a) Precision-recall curve for contact-mode availability prediction. (b) Grasp success rate, comparing generators trained on HUGS and Heur-Multi data.

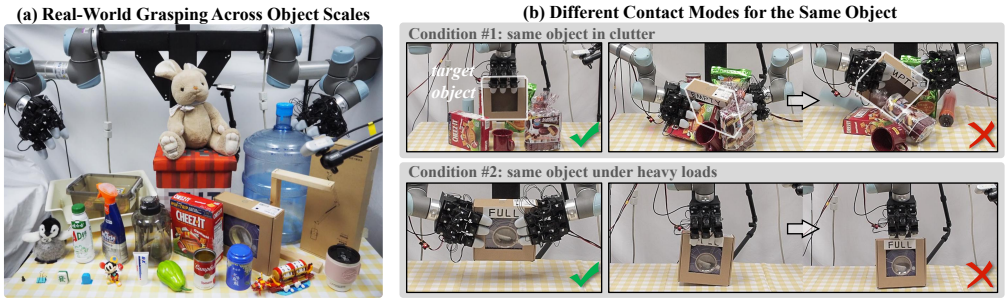


Figure 7: **Real-world grasping demonstrations on LEAP Hands.** (a) Cross-scale grasping over diverse objects. (b) Different mode selection under different deployment constraints.

availability is accurately predicted, reaching a best F1 score of 0.934 in Fig. 6(a). In simulation, the HUGS-trained generator achieves higher grasp success than the Heur-Multi-trained one (Fig. 6(b)), showing that HUGS produces easier-to-learn training data with better tolerance to generated-grasp errors. Success remains above 70% for medium objects (6–13 cm), but drops on very small and large objects, showing that cross-scale, cross-mode grasp generation remains challenging. Since our focus is grasp synthesis, we leave specialized generator design to future work.

4.4 Real-World Cross-Scale Grasping Demonstrations

We demonstrate real-world grasping with LEAP Hands using a generator trained from synthesized LEAP grasps. Across 24 objects, from a screw with a 0.5 cm half-diagonal to a bucket with a 30 cm half-diagonal (Fig. 7(a)), the generator adaptively selects suitable contact modes and grasp poses. Fig. 7(b) shows why mode diversity matters even for the same object: in clutter, a top-down single-hand grasp remains feasible while side bimanual grasping is blocked; when the box is loaded, bimanual grasping becomes necessary to provide sufficient grasping force. Thus diverse generated grasps allow deployment-time constraints to select an appropriate grasp. More grasp cases are shown in Fig. 1 and the appendix; deployment details are also provided in the appendix. Note that the generator is still preliminary, and perception errors from calibration and depth reconstruction remain important bottlenecks for reliable grasping across object scales.

5 Limitations and Conclusion

Limitations. *Human data scale.* HUGS relies on a compact human dataset, so contact-mode and wrist-pose predictions may degrade on far out-of-distribution objects; scaling to large-scale egocentric interaction data is promising but requires further study of mode discovery and annotation. *Online grasp generation.* Our generator is preliminary and remains less reliable than the offline synthesis pipeline, especially for geometrically irregular objects, motivating research on stronger generative models and geometry representations. *Contact Diversity.* The four contact modes in HUGS capture dominant human grasp strategies but do not exhaustively cover the full spectrum of human grasping behaviors [42]. Extending to denser taxonomies or contact maps is needed for functional grasp synthesis beyond stable lifting. *Complete grasping systems.* Our real-world demonstrations remain sensitive to calibration, segmentation, and depth reconstruction errors, while robust deployment also requires end-to-end closed-loop execution, tactile adaptation, and collision-aware motion.

Conclusion. We presented HUGS, a human-prior-guided framework for unified dexterous grasp synthesis across contact modes and object scales. Instead of directly retargeting human demonstrations, HUGS learns an object-conditioned prior over contact modes and wrist poses to guide robot-specific optimization. Experiments show that this prior improves contact-mode prediction, synthesis success, and downstream grasp generation, while real-world demonstrations show adaptive mode selection and grasping from tiny screws to large containers. These results suggest that a human prior learned from compact human grasp data, when used as high-level guidance rather than direct imitation, can support scalable synthesis of diverse and executable dexterous robot grasps across modes and scales.

References

[1] 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. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11359–11366. IEEE, 2023.

[2] J. Chen, Y. Ke, and H. Wang. Bodex: Scalable and efficient robotic dexterous grasp synthesis using bilevel optimization. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 01–08. IEEE, 2025.

[3] Y. Shao and C. Xiao. Bimanual grasp synthesis for dexterous robot hands. *IEEE Robotics and Automation Letters*, 9(12):11377–11384, 2024.

[4] M. Lin, Y.-L. Wei, J. Chen, Y. Lin, S. Chen, J. Lyu, J. Chen, Y. Tang, H. Wang, and W.-S. Zheng. Bidexgrasp: Coordinated bimanual dexterous grasps across object geometries and sizes. *arXiv e-prints*, pages arXiv–2604, 2026.

[5] A. Billard and D. Kragic. Trends and challenges in robot manipulation. *Science*, 364(6446): eaat8414, 2019.

[6] 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*, 40:3897–3916, 2024.

[7] 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 preprint arXiv:2505.05517*, 2025.

[8] S. Wang, Y. Yang, Y. Luo, D. Li, W. Wei, Y. Zhang, P. Hu, Y. Fu, H. Duan, J. Sun, et al. Scaleadfg: Affordance-based dexterous functional grasping via scalable dataset. *IEEE Robotics and Automation Letters*, 2025.

[9] 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. *IEEE Robotics and Automation Letters*, 7(1):470–477, 2021.

[10] R. Zurbrügg, A. Cramariuc, and M. Hutter. Graspqp: Differentiable optimization of force closure for diverse and robust dexterous grasping. In *Conference on Robot Learning*, 2025.

[11] J. Ye, L. Wei, G. Jiang, C. Jing, X. Zou, and X. Wang. From power to precision: Learning fine-grained dexterity for multi-fingered robotic hands. *arXiv preprint arXiv:2511.13710*, 2025.

[12] S. Yang, Y. Xie, Z. Liang, Y. Tian, J. Zeng, D. Lin, and J. Pang. Ultradexgrasp: Learning universal dexterous grasping for bimanual robots with synthetic data. *arXiv preprint arXiv:2603.05312*, 2026.

[13] J. He, D. Li, X. Yu, Z. Qi, W. Zhang, J. Chen, Z. Zhang, Z. Zhang, L. Yi, and H. Wang. Dexvlg: Dexterous vision-language-grasp model at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14248–14258, 2025.

[14] J. Chen, Y. Ke, L. Peng, and H. Wang. Dexonomy: Synthesizing all dexterous grasp types in a grasp taxonomy. *arXiv preprint arXiv:2504.18829*, 2025.

[15] Z.-H. Yin and P. Abbeel. Lightning grasp: High performance procedural grasp synthesis with contact fields. *arXiv preprint arXiv:2511.07418*, 2025.

[16] Y.-L. Wei, J.-J. Jiang, C. Xing, X.-T. Tan, X.-M. Wu, H. Li, M. Cutkosky, and W.-S. Zheng. Grasp as you say: language-guided dexterous grasp generation. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 46881–46907, 2024.

[17] C. Mao, H. Yuan, Z. Huang, C. Xu, K. Ma, and Z. Lu. Universal dexterous functional grasping via demonstration-editing reinforcement learning. *arXiv preprint arXiv:2512.13380*, 2025.

[18] L. Huang, H. Zhang, Z. Wu, S. Christen, and J. Song. Fungrasp: Functional grasping for diverse dexterous hands. *IEEE Robotics and Automation Letters*, 2025.

[19] Y. Ma, K. Chen, K. Zheng, and D. QI. Contact map transfer with conditional diffusion model for generalizable dexterous grasp generation. *Advances in Neural Information Processing Systems*, 38:79807–79836, 2026.

[20] Z. Wu, R. A. Potamias, X. Zhang, Z. Zhang, J. Deng, and S. Luo. Cedex: Cross-embodiment dexterous grasp generation at scale from human-like contact representations. *arXiv preprint arXiv:2509.24661*, 2025.

[21] Z. Wei, Z. Xu, J. Guo, Y. Hou, C. Gao, Z. Cai, J. Luo, and L. Shao. $\mathcal{D}(\mathcal{R}, \mathcal{O})$ grasp: A unified representation of robot and object interaction for cross-embodiment dexterous grasping. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4982–4988, 2025. doi:10.1109/ICRA55743.2025.11127754.

[22] Z. Weng, H. Lu, D. Kragic, and J. Lundell. Dexdiffuser: Generating dexterous grasps with diffusion models. *IEEE Robotics and Automation Letters*, 9(12):11834–11840, 2024.

[23] Y. Xu, W. Wan, J. Zhang, H. Liu, Z. Shan, H. Shen, R. Wang, H. Geng, Y. Weng, J. Chen, et al. Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4737–4746, 2023.

[24] J. Zhang, Z. Ma, T. Wu, Z. Chen, and H. Dong. Cadgrasp: Learning contact and collision aware general dexterous grasping in cluttered scenes. *arXiv preprint arXiv:2601.15039*, 2026.

[25] Y. Zhong, X. Huang, R. Li, C. Zhang, Z. Chen, T. Guan, F. Zeng, K. N. Lui, Y. Ye, Y. Liang, et al. Dexgraspvla: A vision-language-action framework towards general dexterous grasping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 18836–18844, 2026.

[26] J. Zhang, H. Liu, D. Li, X. Yu, H. Geng, Y. Ding, J. Chen, and H. Wang. Dexgraspnet 2.0: Learning generative dexterous grasping in large-scale synthetic cluttered scenes. In *8th Annual Conference on Robot Learning*, 2024.

[27] H. Yuan, Z. Huang, Y. Wang, C. Mao, C. Xu, and Z. Lu. Demograsp: Universal dexterous grasping from a single demonstration. *arXiv preprint arXiv:2509.22149*, 2025.

[28] H. Zhang, Z. Wu, L. Huang, S. Christen, and J. Song. Robustdexgrasp: Robust dexterous grasping of general objects. *arXiv preprint arXiv:2504.05287*, 2025.

[29] Z. Chen, Q. Yan, Y. Chen, T. Wu, J. Zhang, Z. Ding, J. Li, Y. Yang, and H. Dong. Clutterdex-grasp: A sim-to-real system for general dexterous grasping in cluttered scenes. In *Conference on Robot Learning*, pages 885–905. PMLR, 2025.

[30] T. G. W. Lum, M. Matak, V. Makoviychuk, A. Handa, A. Allshire, T. Hermans, N. D. Ratliff, and K. Van Wyk. Dextrah-g: Pixels-to-action dexterous arm-hand grasping with geometric fabrics. In *Conference on Robot Learning*, pages 3182–3211. PMLR, 2025.

[31] R. Singh, A. Allshire, A. Handa, N. Ratliff, and K. Van Wyk. Dextrah-rgb: Visuomotor policies to grasp anything with dexterous hands. *arXiv preprint arXiv:2412.01791*, 2024.

[32] K. M. Lynch and F. C. Park. *Modern robotics*. Cambridge University Press, 2017.

[33] Y.-W. Chao, W. Yang, Y. Xiang, P. Molchanov, A. Handa, J. Tremblay, Y. S. Narang, K. Van Wyk, U. Iqbal, S. Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9044–9053, 2021.

[34] 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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20953–20962, 2022.

[35] W. Cho, J. Lee, M. Yi, M. Kim, T. Woo, D. Kim, T. Ha, H. Lee, J.-H. Ryu, W. Woo, et al. Dense hand-object (ho) graspnet with full grasping taxonomy and dynamics. In *European Conference on Computer Vision*, pages 284–303. Springer, 2024.

[36] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pages 581–600. Springer, 2020.

[37] S. Brahmbhatt, C. Tang, C. D. Twigg, C. C. Kemp, and J. Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *European Conference on Computer Vision*, pages 361–378. Springer, 2020.

[38] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 36(6):1–17, 2017.

[39] C. Choy, J. Gwak, and S. Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019.

[40] E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012.

[41] K. Shaw, A. Agarwal, and D. Pathak. Leap hand: Low-cost, efficient, and anthropomorphic hand for robot learning. *arXiv preprint arXiv:2309.06440*, 2023.

[42] T. Feix, J. Romero, H.-B. Schmiedmayer, A. M. Dollar, and D. Kragic. The grasp taxonomy of human grasp types. *IEEE Transactions on human-machine systems*, 46(1):66–77, 2015.

[43] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, and L. Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024.

[44] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. URL https://arxiv.org/abs/2408.00714.

[45] X. Chen, F.-J. Chu, P. Gleize, K. J. Liang, A. Sax, H. Tang, W. Wang, M. Guo, T. Hardin, X. Li, et al. Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025.

[46] B. Wen, W. Yang, J. Kautz, and S. Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17868–17879, 2024.

[47] R. A. Potamias, J. Zhang, J. Deng, and S. Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12242–12254, 2025.

[48] B. Sundaralingam, S. K. S. Hari, A. Fishman, C. Garrett, K. Van Wyk, V. Blukis, A. Millane, H. Oleynikova, A. Handa, F. Ramos, and N. Ratliff. Curobo: Parallelized collision-free robot motion generation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8112–8119. IEEE, 2023.

[49] B. Tan, C. Sun, X. Qin, H. Adai, Z. Fu, T. Zhou, H. Zhang, et al. Masked depth modeling for spatial perception. *arXiv preprint arXiv:2601.17895*, 2026.

A Implementation Details

A.1 HUGS-Human Dataset Details

Hardware Setup. Raw demonstrations are recorded with four RealSense D435 cameras around the tabletop workspace, as shown in Fig. 8. All cameras are calibrated for both intrinsic and extrinsic parameters, and the tabletop coordinate frame is also calibrated. We record RGB-D streams from all four cameras. The depth images are aligned with the RGB images, and all RGB-D streams from the four camera views are approximately synchronized and captured at a resolution of 1280×720 at 15 Hz. The raw recordings cover the complete manipulation process, including approaching the object, lifting it, and placing it back on the table.

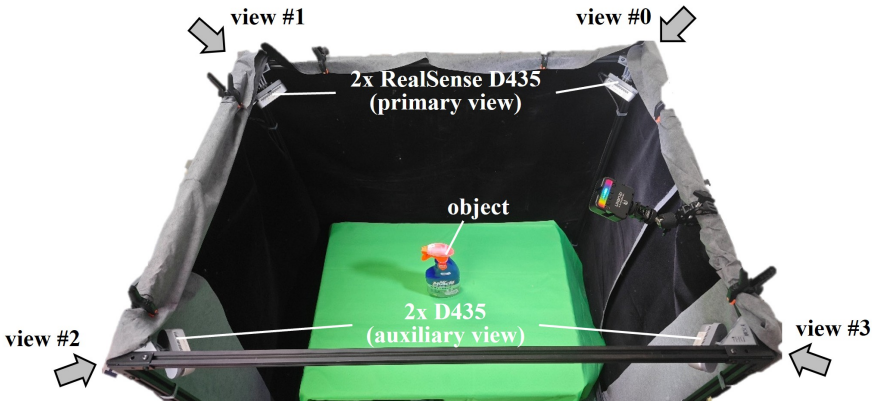


Figure 8: **Hardware setup for collecting the HUGS-Human dataset.** Four RealSense D435 cameras are used for hand-object mesh reconstruction. Camera views #0 and #1 serve as the primary views, while auxiliary views #2 and #3 are used to reduce single-view depth ambiguity for each hand.

Annotation Details. The dataset annotation includes hand-object mesh reconstruction and discrete contact-mode annotation. We use camera views #0 and #1 as the primary views for reconstructing the right and left hands, respectively, and auxiliary views #2 and #3 to reduce single-view depth ambiguity for each hand. View #1 is used for object mesh reconstruction and pose tracking due to its relatively low occlusion. The depth streams are used only to supervise object reconstruction. For each raw recording, we use Grounded-SAM2 [43, 44] for object segmentation, SAM3D [45] for mesh reconstruction, and FoundationPose [46] for pose estimation. The grasp frame is defined as the frame immediately before clear object motion is detected. After detecting the grasp frame, we use WiLoR [47] to estimate MANO hand poses [38] from the primary camera views #0 and #1. To reduce the depth ambiguity inherent in single-view reconstruction, which can lead to hand-object penetration or separation, we triangulate 3D hand keypoints from the primary and auxiliary views using Direct Linear Transform (DLT). We then solve for the rigid transformation that best aligns the WiLoR detections with the triangulated 3D keypoints. Finally, the discrete contact mode c is manually annotated according to the number of hands involved in grasping and the number of dominant contacting fingers, following the four modes defined in Sec. 3. For quality control, we manually inspect all annotations and adjust them when necessary, such as correcting the selected grasp frame. Each data sample includes the object mesh, object pose, hand wrist pose, MANO pose, and contact mode. We will release both raw recordings and processed annotations.

Dataset Visualization. The 304 objects in HUGS-Human are collected from common office and laboratory environments. We visualize the object-scale distribution, per-object grasp count, and contact-mode distribution of HUGS-Human in Fig. 2. We further visualize representative samples from the HUGS-Human dataset at different object scales in Fig. 9, showing both the extracted grasps as reconstructed hand-object meshes and the raw RGB images from all four views. Finally, we

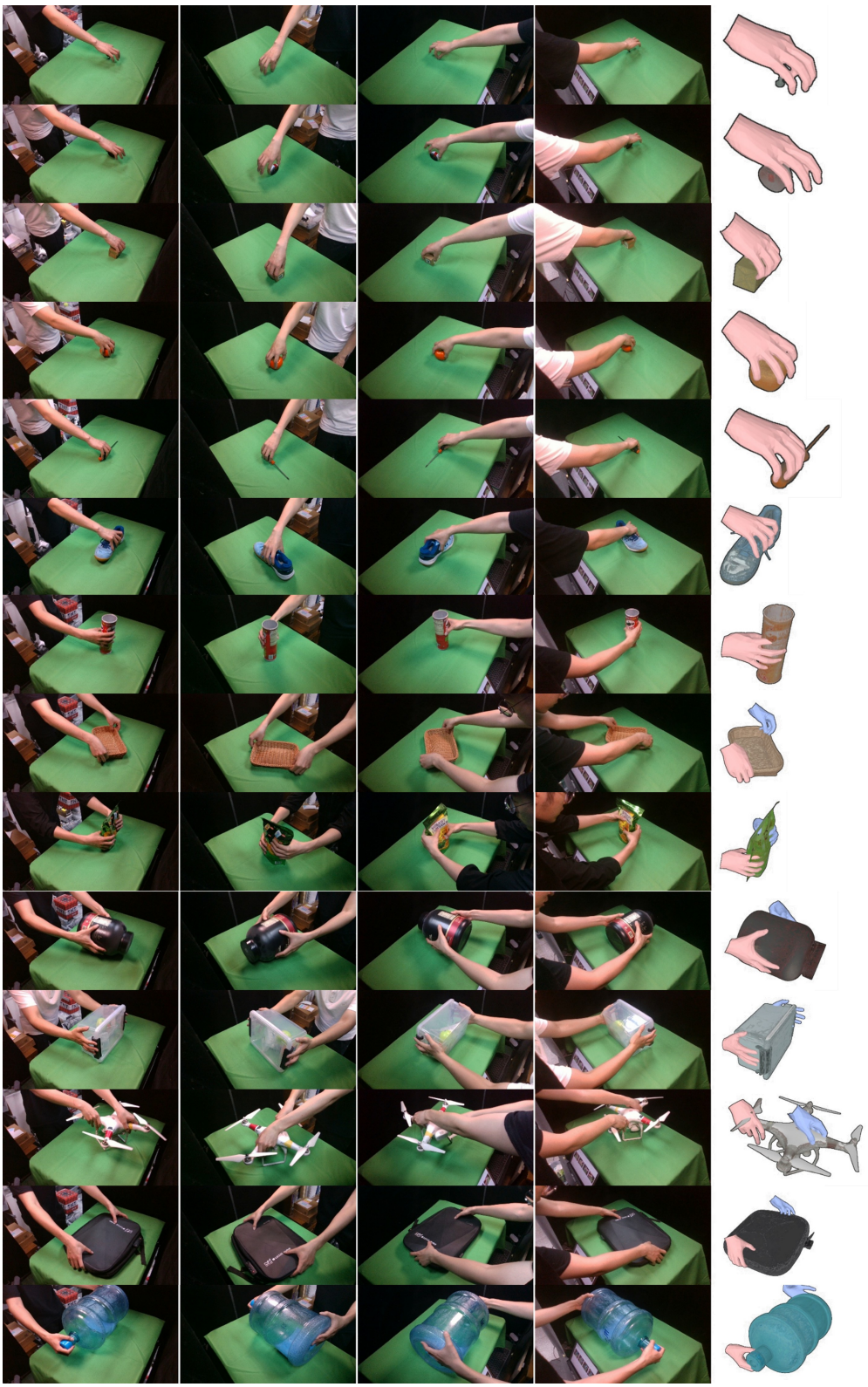


Figure 9: **Visualization of the HUGS-Human dataset.** Representative samples are shown with object scale increasing from top to bottom. Each sample includes RGB images from all four views, along with the rendered hand-object mesh reconstruction.

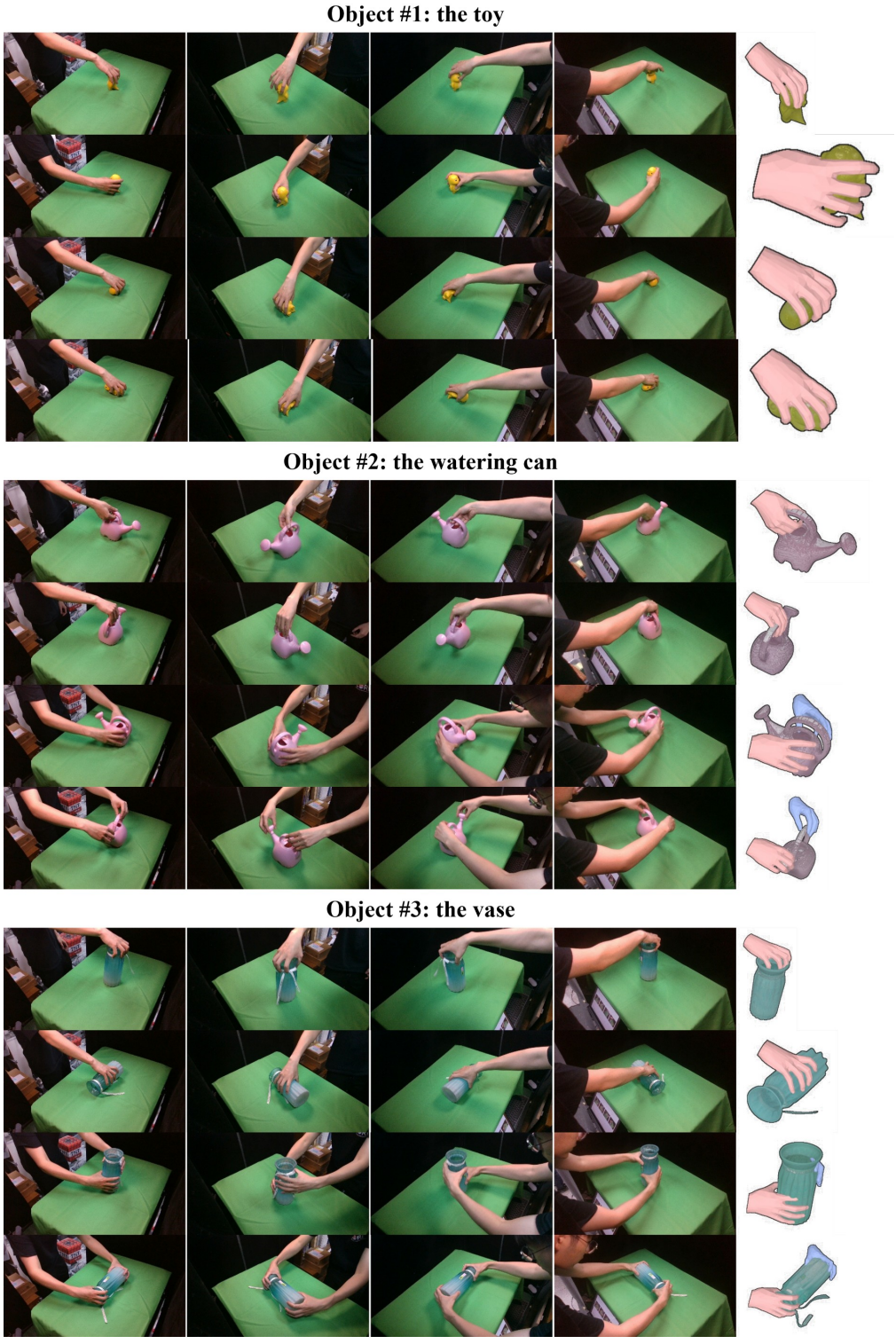


Figure 10: **Visualization of representative cases of multiple grasps collected with a single object.** For the smaller object #1, we collect single-handed grasps with varying numbers of fingers and object poses. For the larger objects #2 and #3, both single-handed and two-handed grasps are collected.

visualize representative cases of multiple grasps collected with a single object in Fig. 10, highlighting the object-level diversity of the HUGS-Human dataset.

Rare Both-Three Contact Mode. At the beginning of data collection, we also planned to include a *Both-Three* contact mode, where two hands are involved but each hand mainly uses sparse three-finger contacts. After completing the collection, however, we found that only a very small number of objects naturally support this mode, resulting in a negligible fraction of the dataset, as shown in Fig. 11. Incorporating *Both-Three* into the HUGS framework is technically feasible, since our contact-mode prior and optimization pipeline can be extended to additional discrete modes. Nevertheless, due to the limited number of valid demonstrations, evaluation results for this mode would be statistically unreliable. We therefore exclude *Both-Three* from the evaluation statistics and focus our analysis on the four dominant contact modes.

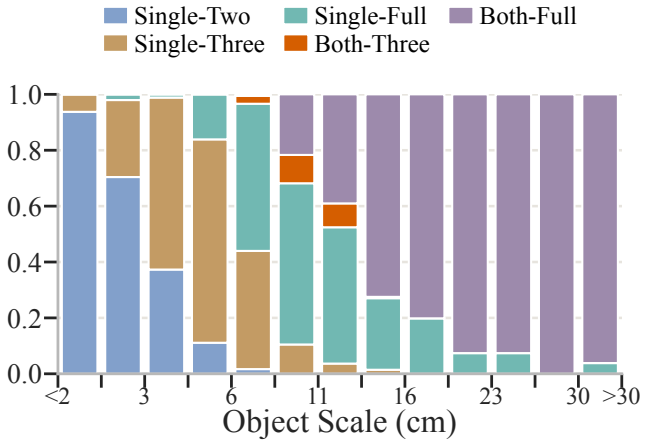


Figure 11: **Contact-mode distribution when including *Both-Three*.** The *Both-Three* mode accounts for only a negligible fraction of the collected demonstrations.

A.2 Human Prior Training Details

Definition of Index MCP Frame. To make the human wrist prior less sensitive to embodiment-specific wrist definitions, we define the predicted pose using an index MCP frame, as shown in Fig. 12. Its translation is the position of the index metacarpophalangeal (MCP) joint, and its rotation follows the dorsal-hand wrist orientation. This representation provides a stable hand-level pose that can be mapped to different robot hands during synthesis.

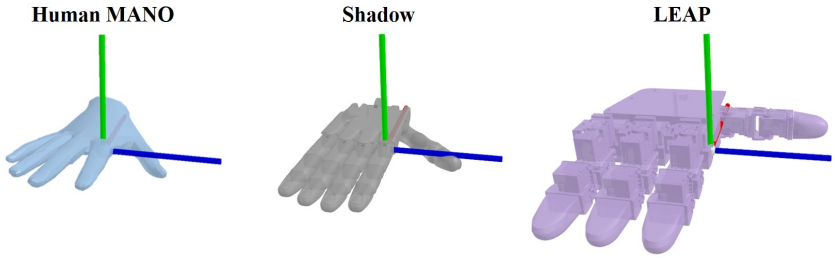


Figure 12: **Definition of the index MCP frame.** We use the index MCP position as the translation anchor and the dorsal-hand wrist orientation as the rotation reference for the human wrist prior.

Architecture and Training. In practice, we train the mode and wrist-pose heads separately because contact-mode prediction converges faster and overfits more easily than wrist-pose generation. Both models take the object point cloud as input. For each object sample, we uniformly sample 1024 points from the object point cloud and center the point cloud before feeding it to the network. The centered

point cloud is sparsely quantized with MinkowskiEngine using a voxel size of 0.005. The point-wise features are aggregated by mean pooling to obtain a 1024-dimensional global object feature. During training, we apply data augmentation including point-cloud centering, random rotations around the z axis, small rotations around the x and y axes with a maximum angle of 3°, and uniform scale augmentation in the range $[0.9, 1.1]$. For scale augmentation, valid hand translation targets are scaled accordingly. We also add Gaussian noise to point coordinates with standard deviation 0.001, clipped to three standard deviations, and apply point dropout with a ratio of 0.1. The contact-mode prior uses a two-layer MLP head. We supervise this branch with posed-object-level soft labels. For multiple human grasp records associated with the same posed object, we compute the occurrence frequency of each contact mode as the target distribution. It is trained with the soft-label cross-entropy loss

$$\mathcal{L}_{\text{mode}} = -\frac{1}{n_{\text{mode}}} \sum_i \sum_k q_{i,k} \log \text{softmax}(\boldsymbol{z}_i)_k, \tag{4}$$

where i indexes the samples included in the mode-supervision set, and k indexes the discrete mode classes. where $\boldsymbol{q}_i$ is the empirical contact-mode distribution computed from all human grasps of the same posed object in the dataset, and $\boldsymbol{z}_i$ denotes the predicted logits. We train this branch from scratch for 100 iterations using AdamW with a batch size of 256. The learning rate is initialized to 10^{-3} and decayed to 10^{-4} with a cosine schedule. The pose prior is a contact-mode-conditioned diffusion model. We use a learnable 128-dimensional embedding for each contact mode and concatenate it with the 1024-dimensional global object feature to form a 1152-dimensional conditional feature. Training samples are uniformly drawn at random from all human grasps in HUGS-Human. The diffusion model generates a normalized 24-dimensional bimanual pose vector $[\boldsymbol{R}_r(9), \boldsymbol{p}_r(3), \boldsymbol{R}_l(9), \boldsymbol{p}_l(3)]$. The translation target $\boldsymbol{p}$ represents the index-MCP position, and the rotation target $\boldsymbol{R}$ uses the 9D representation of the wrist/palm rotation matrix. For right-only grasps, the left hand is filled with a fixed placeholder pose. During training, we apply RMS normalization to the pose vector. We use a cosine noise schedule with 128 training diffusion steps and optimize the denoiser with a velocity-prediction objective. The denoiser takes the noised 24D pose, the 1152-dimensional conditional feature, and the timestep embedding as inputs. It uses hidden layers with Mish activations and outputs a 24D velocity prediction. We train the denoiser with a SmoothL1 denoising objective between the predicted and target velocities. At inference time, we use a 10-step DDIM-style deterministic sampler conditioned on the object feature and contact-mode embedding. We train this branch from scratch for 7500 iterations using AdamW with a batch size of 256. The learning rate is initialized to 10^{-3} and decayed to 10^{-4} with a cosine schedule.

A.3 Grasp Synthesis Details

Contact Regions for Each Contact Mode. Each contact mode specifies the active fingertip regions used by the robot hand during grasp optimization. For *Single-Two*, the active contacts are the thumb and index fingertips of one hand. For *Single-Three*, the active contacts are the thumb, index, and middle fingertips of one hand. For *Single-Full*, the active contacts are all fingertips of one hand. For *Both-Full*, the active contacts are all fingertips of both hands. During grasp optimization, different contact modes only change the expected fingertip contacts involved in the force-closure computation.

Bimanual Optimization Implementation. To adapt bimanual grasp optimization to the frameworks of [2, 48], we model the two hands as a single composite robot by adding dummy joints. The dummy joints include three translational joints and three rotational joints, which parameterize the floating wrist pose of each hand. In practice, very large objects may lie near or beyond the boundary of the human grasp dataset distribution, where training samples are relatively sparse. As a result, the predicted global hand pose can be slightly less accurate, leading to noticeable hand-object penetration in some cases. To make the initial states more suitable for subsequent optimization, we offset each hand by 10 cm along its palm-outward direction when constructing the initial robot wrist poses, as shown in Fig. 21. After sampling wrist poses from the human prior, we first solve batched inverse kinematics to obtain the corresponding dummy joint angles. These IK solutions are then used as the initialization for the subsequent grasp optimization. Unlike the default single-hand configuration in [2], bimanual optimization first optimizes the in-contact grasp pose, and then optimizes the

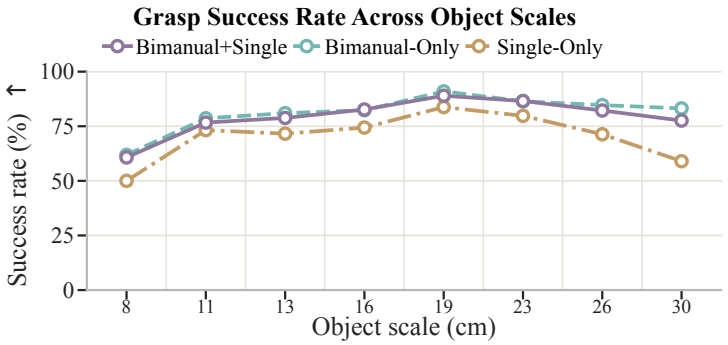


Figure 13: **Ablation of the bimanual force-closure objective.** We compare using only the per-hand force-closure terms (*Single-Only*), only the global bimanual force-closure term (*Bimanual-Only*), and their combination (*Bimanual+Single*). Without considering bimanual force closure, the grasp success rate drops significantly.

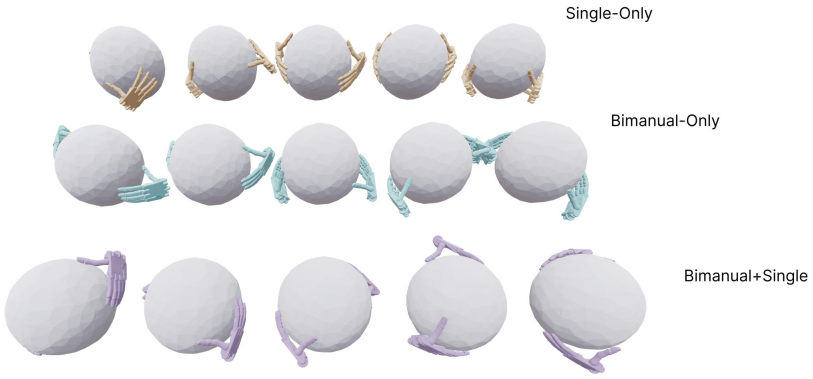


Figure 14: **Qualitative examples of bimanual force-closure ablation.** Using only the global bimanual term (*Bimanual-Only*) can produce degenerate contacts, where each hand touches the object only weakly or at fingertip extremities. Combining global and per-hand terms (*Bimanual+Single*) yields more coordinated bimanual grasps while maintaining better individual hand contact quality.

non-contact pregrasp pose in a final stage by reducing the hand-to-object point distance threshold by 1 cm. This ordering allows the grasp pose to be closer to the object center. In addition, we assign a larger step size to the dummy joint (base pose) variables, which decays over optimization iterations, allowing the wrist poses to adjust more aggressively in early iterations while converging stably. For the force-closure objective, we use a sum of a global bimanual term and per-hand terms: $\Phi_{\text{bi}}(g) = \Phi_{\text{global}}(g) + \Phi_{\text{left}}(g) + \Phi_{\text{right}}(g)$, where Φ_{global} evaluates force closure over contacts from both hands jointly, and Φ_{left}, Φ_{right} evaluate each hand independently. Fig. 13 shows that omitting the global bimanual force-closure term substantially reduces grasp success, indicating that per-hand force closure alone is insufficient for stable bimanual grasping. Although *Bimanual-Only* achieves a success rate similar to the combined objective, Fig. 14 shows that using only the global bimanual term can produce degenerate contacts, where each hand touches the object only weakly or at fingertip extremities. Therefore, the global term is needed for overall bimanual stability, while the per-hand terms improve the contact quality of each individual hand.

Human-to-Robot Scale Ratio. Before querying the human prior, we rescale the object point cloud according to the human-to-robot hand scale ratio. The motivation is to match the relative object size perceived by the robot hand to that in the human demonstrations, since different hand embodiments have different physical sizes as illustrated in Fig. 12. For example, the LEAP Hand is larger than a human hand, so an object grasped by a human appears relatively smaller to LEAP. Therefore, when

querying the human prior for LEAP synthesis, we downscale the object point cloud before prediction. We use a scale ratio of 1.0 for the Shadow Hand and 1.4 for the LEAP Hand.

MuJoCo Simulation. We filter optimized grasps in MuJoCo simulation by lifting the object and checking three criteria for success. First, the object must be lifted to a height of 0.1 m. Second, the object pose must remain stable throughout the lift, with position deviation below 5 cm and rotation deviation below 15°. Third, the grasp must be collision-free. A grasp is disqualified if the hand is in self-collision or intersects the object at the start of the simulation, or if self-collision occurs at any point during the grasping process. For physical simulation across object scales, we assign a scale-dependent effective density to each object. To avoid cubic mass growth under geometric scaling, we use $d(s) = d_0(s/s_0)^{-2}$, where $d_0 = 700\,\mathrm{kg/m^3}$ and $s_0 = 0.06\,\mathrm{m}$. This formulation makes the resulting object mass scale approximately linearly with s. The resulting median object masses are 0.032 kg at scale 0.02, 0.096 kg at scale 0.06, 0.160 kg at scale 0.1, and 0.339 kg at scale 0.3.

A.4 Grasp Generator Details

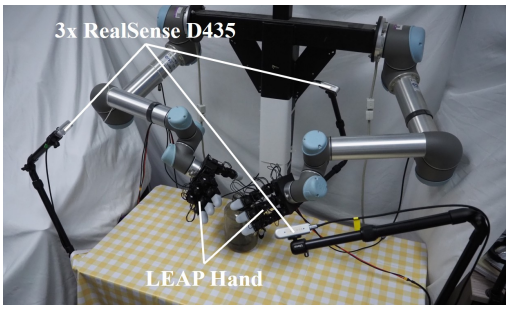
Architecture. The robot grasp prior uses the same sparse point-cloud encoder as the human prior. The input partial object point cloud is fused from three camera views. We sample 1024 points from the partial object point cloud, apply MinkowskiEngine sparse quantization, and feed the quantized point cloud into MinkUNet. Mean pooling over point-wise features produces a 1024-dimensional object feature. The model contains a contact-mode availability head and a mode-conditioned robot pose head. The availability head predicts independent binary availability scores for the four contact modes. Its supervision is computed from the contact-mode records of the same scene in the synthetic robot dataset, allowing multiple contact modes to be available simultaneously. The pose-head architecture follows [14]. It conditions on the concatenation of the object feature and a learnable 128-dimensional contact-mode embedding. It first uses a conditional diffusion model to generate the final grasp-frame 24-dimensional bimanual wrist pose $[\boldsymbol{R}_r, \boldsymbol{p}_r, \boldsymbol{R}_l, \boldsymbol{p}_l]$, where rotations use the 9D rotation-matrix representation. The diffusion model uses RMS-normalized poses, 128 training diffusion steps, a cosine noise schedule, and a velocity-prediction objective, and is trained with a SmoothL1 denoising loss. To recover the full robot trajectory, we further use two MLPs for the left and right hands. Conditioned on the final wrist pose, these MLPs regress wrist translations and rotations for the pregrasp and grasp stages, as well as finger joint positions for the pregrasp, grasp, and squeeze stages.

Training. The overall training objective consists of five terms: contact-mode availability BCE loss, final-wrist diffusion denoising loss, trajectory translation loss, trajectory rotation loss, and joint-position loss. We use equal weights for all five terms in the current configuration and train the entire network jointly in a single stage for 50000 iterations. We use AdamW with a batch size of 256. The learning rate is initialized to 10^{-3} and decayed to 10^{-4} with a cosine schedule.

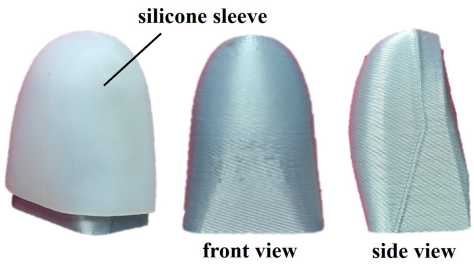
Sampling. At inference time, the availability head first predicts independent contact-mode scores for the input partial point cloud. We select feasible contact modes using a score threshold of 0.9. According to the precision-recall curve in Fig. 6(a), this threshold yields approximately 95% precision and 90% recall for the Shadow Hand. For each selected contact mode, we sample the final wrist pose with the DDIM deterministic sampler conditioned on the object feature and contact-mode embedding, followed by RMS de-normalization. The trajectory MLPs then predict the pregrasp and grasp wrist poses, as well as the pregrasp, grasp, and squeeze joint positions, conditioned on the sampled final wrist pose. For each feasible contact mode, we sample 100 grasp poses and select the top 10 candidates according to the approximate diffusion probability for simulation evaluation. The resulting candidates may be further filtered by reachability, collision checking, and task-specific deployment constraints for downstream applications.

A.5 Real-World Deployment Details

Deployment Pipeline. The real-world hardware setup is shown in Fig. 15. We deploy the system on a dual-arm UR5 platform with two LEAP Hands equipped with customized fingertips. During deployment, three RealSense D435 cameras capture RGB-D observations. All cameras are calibrated



(a) Real-world hardware setup



(b) Customized LEAP fingertip

Figure 15: **Real-world hardware setup and customized LEAP fingertip.** Our hardware setup consists of three RealSense D435 cameras arranged approximately 120° apart for object point-cloud reconstruction. Compared with the original LEAP fingertip, our customized fingertip has a smoother geometry and is covered with a silicone sleeve to improve contact stability.

for inter-camera extrinsics, and we further perform hand-eye calibration between the robot and the primary camera. RGB images are segmented with Grounded-SAM2 [43, 44], depth maps are refined with Lingbot-depth [49], and the fused object point cloud is fed to the generator. Collision-free approach motions are planned using cuRobo [48]. For real-world execution, we sample 25 grasps for each feasible contact mode and manually select three candidates for robot execution. Snapshots of the executed real-world grasps on objects #1–#23 are shown in Fig. 22 and Fig. 23, while object #24 is used to demonstrate the significance of contact-mode diversity (Fig. 7 (b)). As discussed in Sec. 5, real-world generation is affected by calibration errors, depth reconstruction noise, segmentation quality, and other deployment-specific factors. Moreover, our online generator is still preliminary and remains less reliable than the offline synthesis pipeline. Therefore, the real-world experiments are intended to demonstrate the deployment potential of HUGS in physical scenes, rather than to provide a quantitative evaluation of grasp success rate. The typical failure modes are analyzed in Appendix D.3. Building a fully autonomous and robust real-world grasping system remains an important direction for future work.

B Evaluation Details

B.1 Metrics Details

Human Contact-Mode Distribution Prediction. This section provides the metric definitions for the contact-mode distribution prediction experiment in Sec. 4. Given a posed object, the task is to predict the empirical distribution of human-preferred contact modes rather than a single contact-mode label. Let $\boldsymbol{q}$ denote the empirical contact-mode distribution and $\boldsymbol{p}$ denote the predicted distribution over contact modes. We measure the discrepancy between the two distributions using

$$D_{\mathrm{KL}}(\boldsymbol{q}\|\boldsymbol{p}) = \sum_t q_t \log \frac{q_t}{p_t}. \tag{5}$$

We also report soft precision,

$$\mathrm{SoftPrec}(\boldsymbol{p}, \boldsymbol{q}) = \sum_{t \in \mathcal{P}(\boldsymbol{q})} p_t, \quad \mathcal{P}(\boldsymbol{q}) = \{t \mid q_t > 0\}, \tag{6}$$

which measures how much predicted probability mass falls on ground-truth positive contact modes. Finally, we report soft recall,

$$\mathrm{SoftRec}(\boldsymbol{p}, \boldsymbol{q}) = \sum_{t \in \mathrm{Top}_{|\mathcal{P}(\boldsymbol{q})|}(\boldsymbol{p})} q_t, \tag{7}$$

which measures how much ground-truth probability mass is covered by the top predicted modes, with the number of selected modes set to $|\mathcal{P}(\boldsymbol{q})|$, i.e., the number of ground-truth positive contact modes. Lower KL and higher soft precision/recall indicate better contact-mode prediction.

Synthesis Success Rate. For a given evaluation group, the synthesis success rate is defined as the fraction of optimization attempts that pass physical validation in simulation, i.e., $\mathrm{SuccessRate} = N_{\mathrm{success}}/N_{\mathrm{attempt}}$, where N_{attempt} is the number of grasp optimization attempts and N_{success} is the number of physically successful grasps.

Success Counts per Scene. For each object scene, the success count is the number of successful grasps obtained from all optimization attempts allocated to that scene. When reporting success counts per scene across an object-scale bin or a method, we average this count over all scenes in the corresponding group.

Pose Diversity. We measure grasp-pose diversity using the explained-variance ratio of the first principal component, following the PCA-based diversity metric in prior work [2, 14]. For each synthesized grasp, we use the optimized wrist pose and hand joint angles as the pose feature, including wrist translation, wrist rotation, and finger joint positions. For each object scene, we perform PCA over all validated grasp poses of that scene and record the explained-variance ratio of the first principal component. We then average this value over all scenes in the corresponding evaluation group. This differs from [2, 14], which compute PCA after pooling grasps across scenes. We adopt the scene-wise formulation because our focus is the diversity of valid grasp poses for the same scene. A lower first-component explained-variance ratio indicates that successful grasps are less dominated by a single pose direction and are therefore less concentrated along one major mode. In our analysis, however, large dispersion can also arise from unnatural wrist poses produced by heuristic sampling; thus the diversity metric should be interpreted together with grasp naturalness and qualitative examples.

Robot Contact-Mode Availability Prediction. For the robot grasp generator, contact-mode availability is evaluated as a multi-label binary prediction problem. For each scene and contact mode, the ground-truth label is positive if the synthesized dataset contains at least one validated grasp of that mode for the scene, and negative otherwise. Given predicted availability scores, we sweep the decision threshold to compute precision and recall: $\mathrm{Precision} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$ and $\mathrm{Recall} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$. The F1 score is defined as $\mathrm{F1} = 2\,\mathrm{Precision}\,\mathrm{Recall}/(\mathrm{Precision}+\mathrm{Recall})$. We report the precision-recall curve and the best F1 score over thresholds.

B.2 Baseline Details

Scalar-Scale Rules. The scale-rule baseline is derived from statistics of HUGS-Human. Here, object scale refers to the half-diagonal length of the object's axis-aligned bounding box (AABB). We first group posed objects in the training split into scale bins according to their object size. For each scale bin, we aggregate all human grasps associated with posed objects in that bin and compute the empirical contact-mode distribution. At test time, a posed object is assigned to the corresponding scale bin based only on its scalar object size, and the precomputed bin-level contact-mode distribution is used as the prediction. This baseline therefore captures scale-dependent mode preferences but ignores detailed object geometry. In the contact-mode distribution prediction experiment in Sec. 4, we evaluate the continuous contact-mode distributions directly. For grasp synthesis, however, the scalar-scale rules are used only to decide whether each contact mode should be considered. All contact modes selected by the rule are assigned the same optimization budget.

Heuristic Wrist-Pose Sampling. For heuristic baselines, we use convex-hull wrist sampling to initialize optimization. For each object scene, we first construct an expanded convex hull of the object. After applying the object scale in the scene, we expand the hull along the hull-vertex normals. We then uniformly sample candidate points on the expanded hull surface and transform both the sampled points and face normals into the world frame. Each candidate point defines a palm seed. The translation is set to the sampled point, and the approach axis is set opposite to the outward normal so that the palm faces the object. Candidate seeds are filtered by collision-free checks. For bimanual grasping, we first randomly sample a right-hand surface candidate. We then compute the mean xy position of all valid candidates in the current scene as the pairing center, mirror the right-hand xy position about this center, and keep the right-hand z coordinate unchanged. In the left-hand candidate

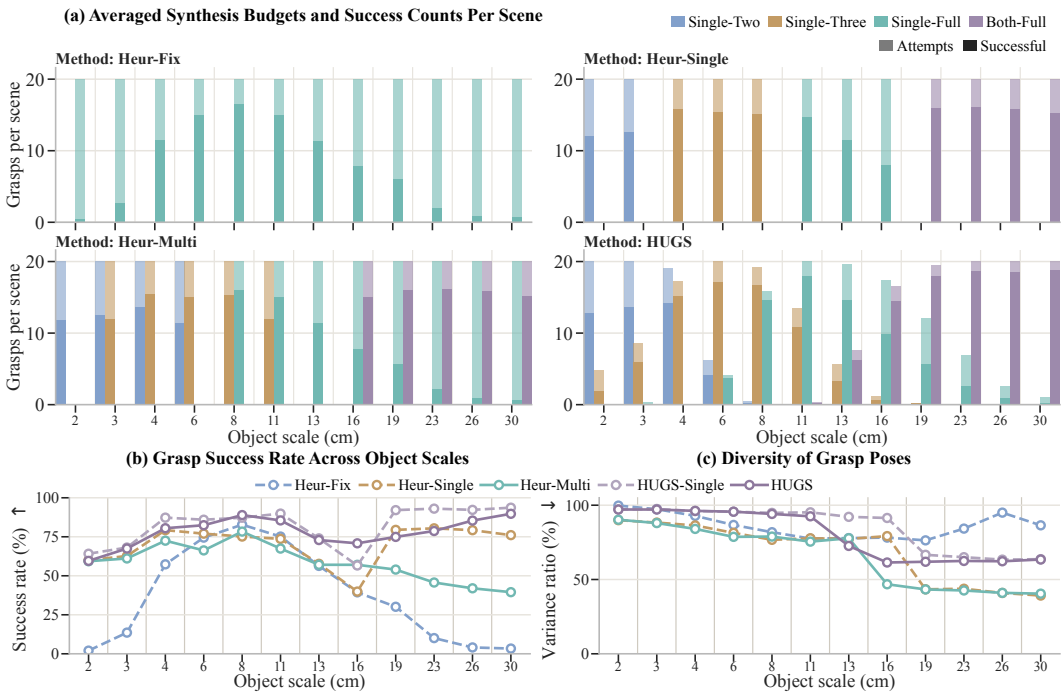


Figure 16: **Contact-mode allocation and synthesis success on the LEAP Hand.** (a) Averaged synthesis budgets and success counts per scene. (b) Overall synthesis success rate across object scales. (c) Pose diversity measured by the explained-variance ratio of the first principal component.

set, we select the real surface candidate whose translation is closest to this mirrored target. This procedure only pairs translations; the left-hand rotation is still constructed from the surface normal of the selected left-hand candidate, rather than by mirroring the right-hand orientation. Finally, we add translational and rotational perturbations to the sampled poses and convert them into initial poses.

C Additional Results

C.1 LEAP Hand Simulation Results

We provide additional LEAP Hand simulation results using the same synthesis protocol as the Shadow Hand experiments, except that we use a human-to-robot scale ratio of 1.4 when querying the human prior. Since the LEAP Hand is larger than a human hand, an object grasped by a human appears relatively smaller to LEAP; therefore, we downscale the object point cloud before prediction.

LEAP Synthesis Results. The LEAP Hand results in Fig. 16 show trends consistent with the Shadow Hand results in Sec. 4. Across object scales, HUGS adaptively reallocates synthesis budget across contact modes instead of relying on fixed scalar-scale rules. This object-conditioned allocation improves synthesis success over heuristic baselines while preserving multi-mode grasp coverage. The diversity trend is also consistent with the Shadow Hand experiments: HUGS concentrates optimization around human-preferred wrist regions, whereas heuristic sampling can introduce larger pose dispersion through less natural wrist initializations. These results suggest that the learned human prior is not tied to a single robot hand and can transfer to a different dexterous hand, even when its physical size differs from that of a human hand.

Qualitative Synthetic LEAP Grasps. Fig. 17 shows qualitative examples of HUGS-synthesized grasps on the LEAP Hand. The examples span different object scales and contact modes, from sparse single-hand grasps on small objects to full-hand and bimanual grasps on larger objects. They also include cases where different contact modes are synthesized for the same object.

Figure 17: **Synthesized LEAP Hand grasps across object scales and contact modes.**

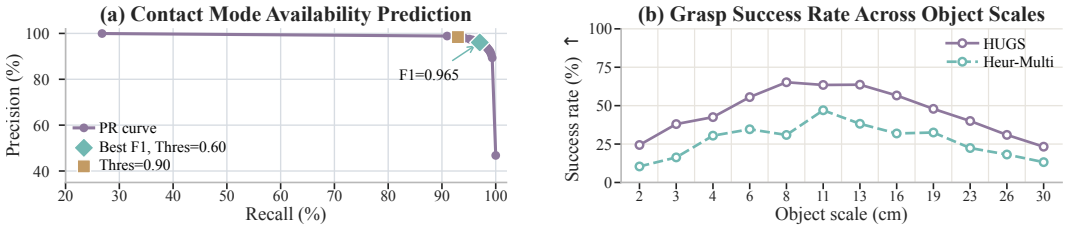


Figure 18: **Distilling from synthesized LEAP grasps.** (a) Precision-recall curve for contact-mode availability prediction. (b) Grasp success rate, comparing generators trained on HUGS and Heur-Multi LEAP grasp data.

LEAP Distillation Results. Fig. 18 further evaluates distillation from synthesized LEAP grasps, following the same protocol as the Shadow Hand generator experiment in Fig. 6. The results are consistent with the Shadow Hand setting: the generator trained on HUGS data predicts contact-mode availability and produces more successful grasps than the generator trained on heuristic data. This indicates that the synthesized HUGS grasps remain useful for supervising online grasp generators after transferring the synthesis pipeline to the LEAP Hand.

C.2 More Visualization

Contact-Mode Prior Visualization. Fig. 19 visualizes contact-mode distributions predicted by the learned human prior for diverse object point clouds. The examples show that the prior adapts its mode probabilities according to object geometry and scale, assigning high probability to sparse single-hand modes for small objects and to full-hand or bimanual modes for larger objects.

Wrist-Pose Prior Visualization. Fig. 20 shows mode-conditioned wrist-pose samples from the human prior. For the same object, different contact modes can induce different preferred approach regions. In Case 1, the object is large and wide. Bimanual grasps can approach either from the side or from above because the two hands can jointly cover the object, whereas single-hand side grasps would place the contact region far from the object's center of mass. Therefore, the single-hand prior favors top-down grasps whose contact region is closer to the gravity line through the center of mass. In Case 2, the object is approximately cylindrical with height larger than its radius. From a top-down approach, the small radius leaves limited space around the object, making two-finger grasps more suitable; from an oblique side approach, the object height provides enough accessible surface for three-finger contacts, so the *Single-Three* prior favors oblique wrist poses.

From Human Prior to Robot Grasps. Fig. 21 visualizes how the learned human prior guides robot grasp synthesis. The sampled human-prior wrist poses provide coarse object-conditioned guidance, which is transferred to robot wrist initializations and then refined by force-closure-aware optimization. The examples cover both single-hand and bimanual grasps across different object scales.

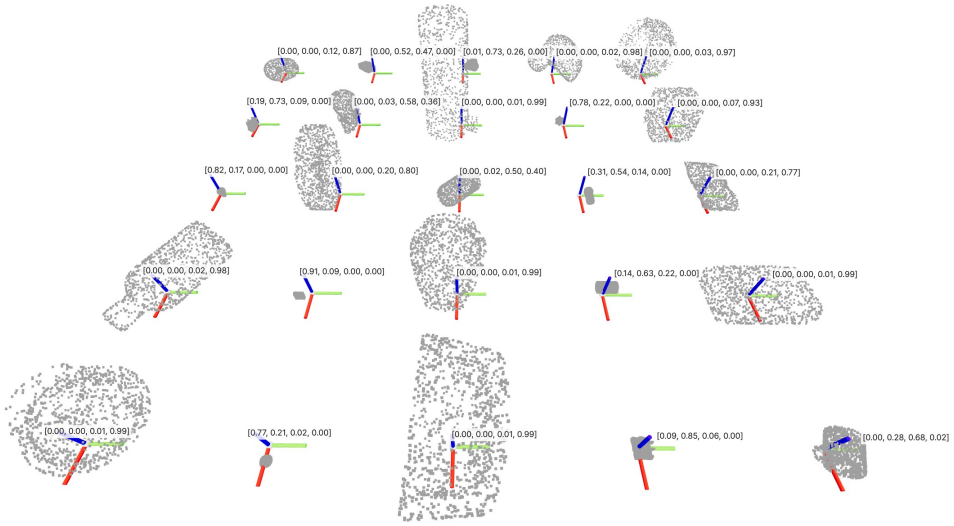


Figure 19: **Predicted contact-mode distributions from the human prior.** Each example shows an object point cloud together with the predicted probability vector over contact modes, ordered as [*Single-Two*, *Single-Three*, *Single-Full*, *Both-Full*]. The coordinate axes have a length of 0.1 m.

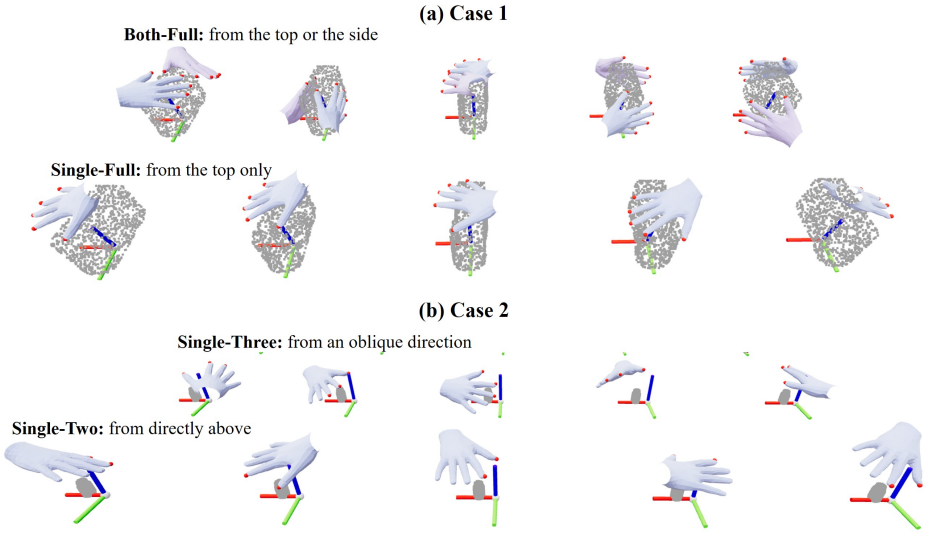


Figure 20: **Object-conditioned wrist-pose samples for different contact modes.** The examples show that the learned pose prior predicts wrist-pose distributions conditioned jointly on object geometry and contact mode. For a given object geometry, different contact modes favor different approach regions: in Case 1, *Both-Full* admits both top and side bimanual approaches, while *Single-Full* mainly concentrates on top-down approaches; in Case 2, *Single-Three* favors oblique approaches, whereas *Single-Two* favors more direct top-down approaches.

Real-World Grasping Snapshots. We provide real-world grasping snapshots in Fig. 22 and Fig. 23. The objects cover a wide range of scales and are approximately ordered from smaller to larger instances. Depending on the object geometry, each object may support one or two contact modes. For each feasible contact mode, we execute three distinct grasps to illustrate the diversity of generated bimanual grasp behaviors.

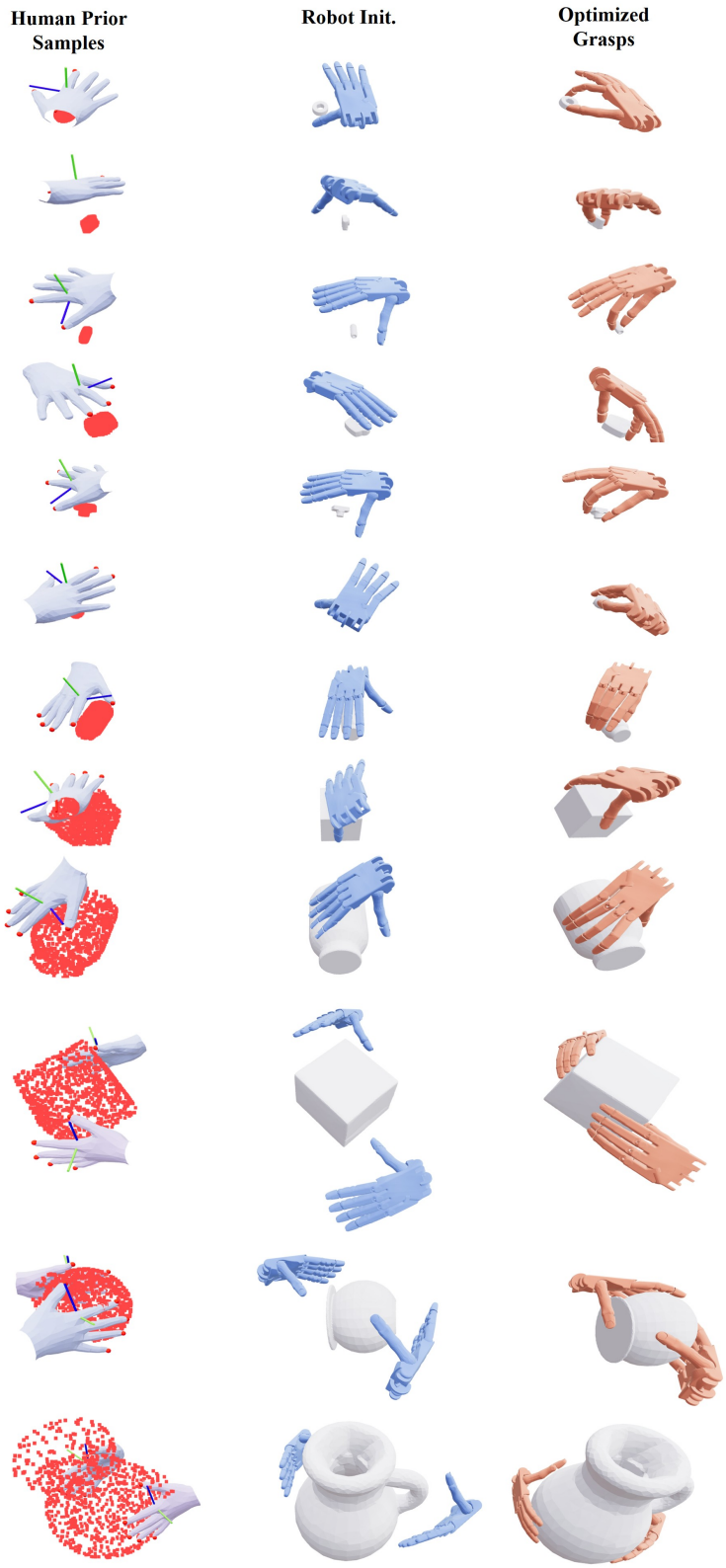


Figure 21: **From human prior samples to optimized robot grasps.** Left: contact-mode and wrist-pose samples predicted by the human prior for each object. Middle: robot wrist initializations obtained by transferring the prior samples to the robot hand. Right: final robot grasps after force-closure-aware optimization.

Figure 22: **Snapshots of real-world grasp on objects #1–#12.** Objects are approximately ordered by increasing scale from top to bottom. Some objects allow only one feasible contact mode, whereas others allow two. For each feasible contact mode, we execute three distinct grasps.

Figure 23: **Snapshots of real-world grasp on objects #13–#23.**

D Failure Analysis

D.1 Failure Analysis of Human-Prior-Guided Grasp Synthesis

Fig. 24 summarizes typical failure modes observed during human-prior-guided grasp synthesis. Although the learned prior provides useful global guidance, it remains a coarse prior. First, the human prior and the optimization objective do not explicitly model functional grasp awareness, so the pipeline may produce grasps that are geometrically plausible but functionally weak. Second, the contact-mode prior can sometimes predict *Single-Full* for objects that are slightly beyond the capability of single-hand grasping. Third, for relatively small objects, the *Both-Full* prior can produce partially overlapping hands. Finally, for extremely large objects, the *Both-Full* prior can produce noticeable hand-object penetration. As described in Sec. A.3, the bimanual initialization offset helps reduce such problematic initial states. These hand-hand overlap and hand-object penetration cases can be further mitigated by the feasibility-constrained physical optimization stage. This highlights the benefit of our human-prior-plus-robot-optimization design: the learned human prior provides diverse and global grasp hypotheses, while the robot optimization stage explicitly enforces physical plausibility and adapts the grasps to the target robot embodiment. These failure modes reveal limitations of the current prior and highlight the need for larger and more comprehensive human grasp datasets, together with stronger functional reasoning for grasp synthesis.

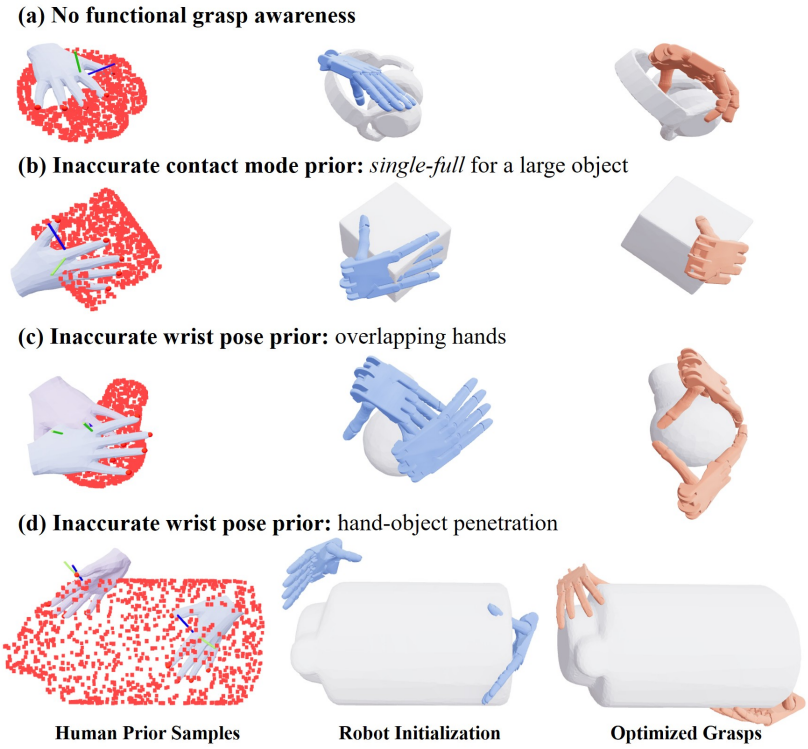


Figure 24: **Failure analysis of human-prior-guided grasp synthesis.** Each row shows a failure mode from human prior samples to robot initialization and optimized grasps. (a) The human prior and the optimization objective lack functional grasp awareness, so the pipeline may produce geometrically plausible but functionally weak grasps. (b) The contact-mode prior can sometimes predict *Single-Full* for objects that are slightly beyond the capability of single-hand grasping. (c) For relatively small objects, the *Both-Full* prior can sometimes produce partially overlapping hands. (d) For extremely large objects, the *Both-Full* prior can sometimes produce clear hand-object penetration. The hand-hand overlap and hand-object penetration cases can often be mitigated during optimization.

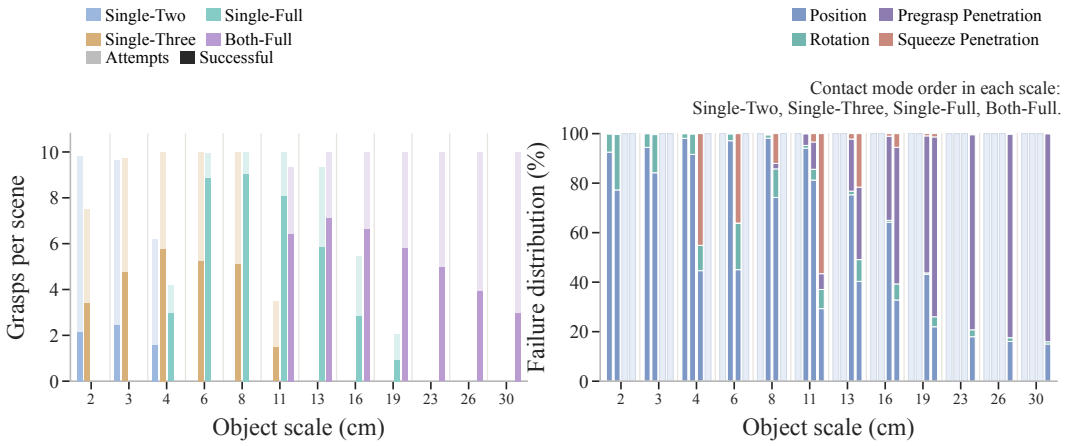


Figure 25: **Failure distribution of the distilled grasp generator on the Shadow Hand.** Left: number of generated and successful grasps per scene across object scales and contact modes. Right: distribution of failure causes for each contact mode at each object scale, ordered as *Single-Two*, *Single-Three*, *Single-Full*, and *Both-Full*. Failures are categorized as position error (unsuccessful lifting), rotation error (excessive object rotation), pregrasp penetration, and squeeze-stage self-penetration.

D.2 Failure Analysis of Distilled Robot Grasp Generators.

Fig. 25 further analyzes generator failures after distillation by reporting both the simulation success rates of generated grasps and the distribution of failure causes across object scales and contact modes. The results reveal several scale- and mode-dependent failure patterns. First, most failures are caused by position error, where the generated grasp does not successfully lift the object. This indicates a remaining limitation in the accuracy of grasp-pose generation. For Single-Full grasps, a noticeable portion of failures comes from squeeze-stage self-penetration, which is not explicitly considered in the current squeeze-pose synthesis process [2]. This issue is particularly pronounced on the Shadow Hand, whose relatively small inter-finger spacing makes self-penetration more likely during finger closure. We plan to address this limitation in future work. For large objects, especially under the Both-Full mode, many failures already occur at the pregrasp stage due to hand-object penetration, suggesting that the generated global poses are not sufficiently accurate for large-scale objects. These observations indicate that, although HUGS-synthesized data can effectively supervise online generators, designing robust robot grasp generators across scales and contact modes remains an important direction for future research.

D.3 Failure Analysis of Real-World Grasping

In addition to the generative-network limitations analyzed in the simulation evaluation, real-world failures are affected by several deployment-specific factors, as illustrated in Fig. 26. First, low-quality grasp predictions can produce loose grasps that fail to securely lift the object. Second, calibration, segmentation, and depth reconstruction errors can corrupt the reconstructed object point cloud, causing the generated grasp to miss the target object, especially for small objects. Third, geometrically irregular objects and challenging cross-scale cases may lie outside the training distribution, leading to degraded grasp generation. These observations highlight the need for advances at both the model and system levels. On the model side, future work should improve the accuracy and generalization of grasp generators to more diverse object distributions. On the system side, robust real-world deployment will require end-to-end closed-loop execution, tactile adaptation, and collision-aware motion planning.

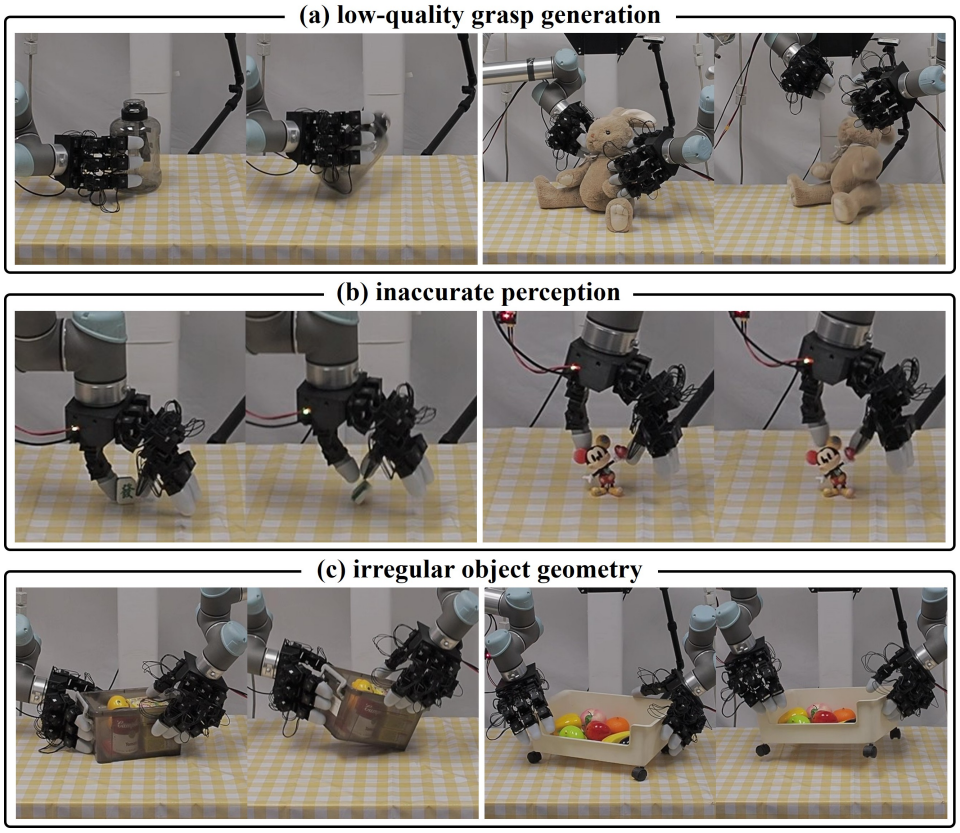


Figure 26: **Typical failure modes of real-world grasping.** (a) Low-quality grasp generation may produce loose or unstable grasps, suggesting that the current grasp generation network can be further improved. (b) Inaccurate calibration, segmentation, and depth sensing reduce the reliability of reconstructed point clouds, resulting in grasps that miss the object. This issue is particularly pronounced for small objects. (c) Irregularly shaped objects often suffer from poor point-cloud reconstruction and may lie outside the distribution of the training object set, leading to degraded grasp generation. This suggests that the diversity and coverage of the synthetic object dataset can be further improved to enhance generalization.