SynManDex: From Human Pre-Grasps to Executable Dexterous Robot Grasps

Yanming Shao[1] Zanxin Chen[1,2] Wenwei Lin[3] Mingjie Zhou[4] Tianxing Chen[5] Xiaokang Yang[2] Yichen Chi[6,†]

Yao Mu[2,†]

[1]Shanghai AI Lab [2]SJTU [3]Shenzhen University [4]Fudan University [5]University of Hong Kong [6]ZTE Corporation

[†]Corresponding authors. Project page: https://tsunami-kun.github.io/SynManDex/

Abstract

Dexterous robot grasping requires both task-aware contact choices and embodiment-specific physical validity. Human hand-object data provides strong functional priors, but direct MANO-to-robot transfer often produces penetration, missed contacts, or unreachable wrist poses; robot-native force-closure optimizers satisfy contact mechanics but can converge to semantically unnatural grasps. We introduce SynManDex, a synthetic data-generation pipeline that uses generated human pre-grasps only as pre-contact proposals and delegates contact formation to robot-native optimization. Given an object mesh, an object-conditioned diffusion model samples MANO pre-grasps that encode approach direction, hand role, and coarse finger coordination. These seeds are retargeted to the target hand, refined with collision and force-closure objectives, and filtered by arm-hand IK and dynamic lift rollouts, yielding executable bimanual grasp-and-lift demonstrations. On 312 objects from 25 classes, SynManDex achieves 86.4% force closure with 0.6 mm penetration and a 4.67/5 combined human-likeness audit score. Policies trained on the admitted trajectories reach 80.7% success on held-out simulated objects and 25/30 successful zero-shot hardware trials on a three-object benchmark. A Shadow-hand diagnostic further suggests that human-prior seeding can improve contact-basin discovery for a different dexterous morphology.

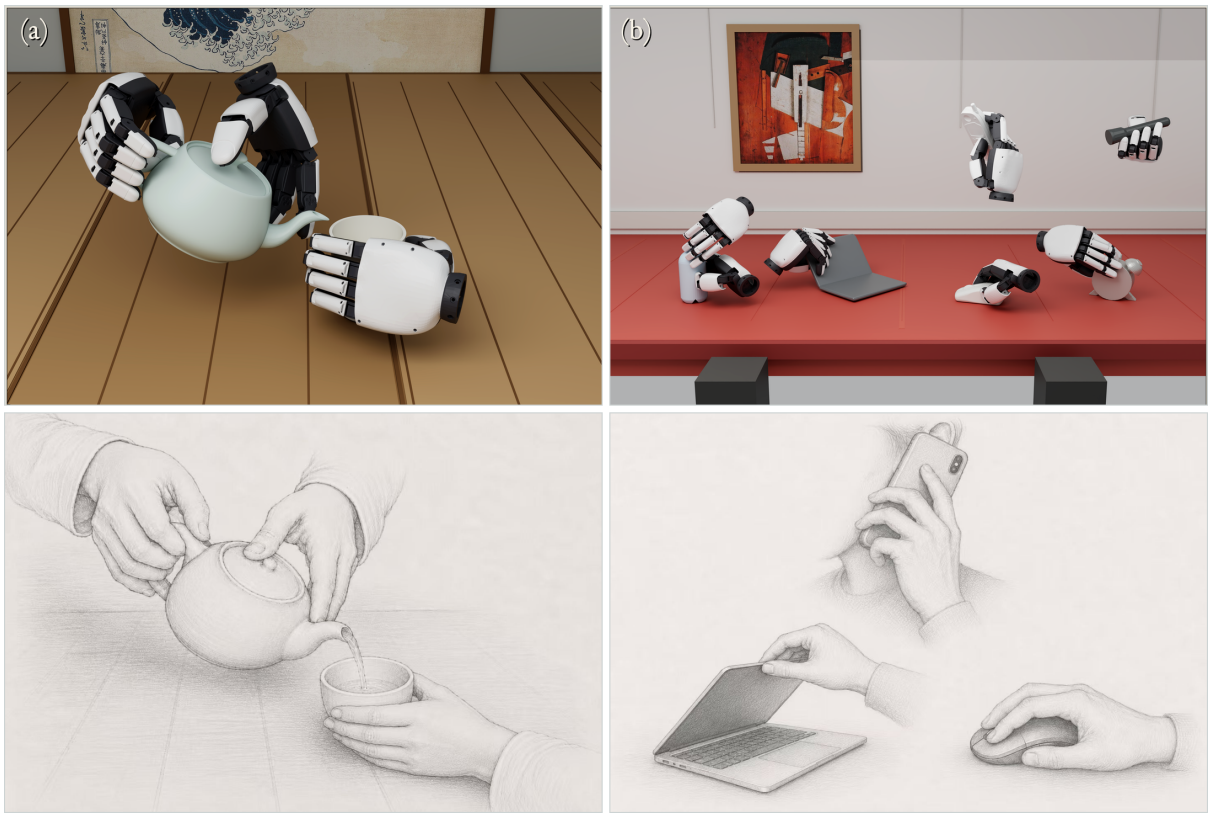

Figure 1. SynManDex uses human-like hand priors to synthesize dexterous robot grasps. Top row: XHand teaser panels (a,b) show robot-native optimized grasps that preserve human-like hand roles while satisfying physical execution constraints. Bottom row: human sketches illustrate the corresponding functional hand-object priors for collaborative tea pouring and office-object interaction.

1 Introduction

Dexterous grasping is difficult because a successful grasp is both semantic and mechanical: the fingers should contact functionally meaningful object regions, while the resulting contact set must be reachable, collision-free, and stable for the target robot hand. This coupling is especially sharp for bimanual systems, where wrist poses, high-DoF finger configurations, inter-hand collision, and gravity all constrain which human-like contacts can become executable robot actions [1, 2].

Existing methods tend to emphasize one side of this tradeoff. Physics-first grasp synthesis can optimize contact and stability, but it does not explicitly encode human/task intent [3–6]. Human-prior, taxonomy, and vision-language methods provide useful cues about approach direction, affordance, and grasp type [7–10], but their outputs are often not directly executable by a target robot hand. Direct retargeting appears natural, yet morphology mismatch can reintroduce missed contacts, interpenetration, and unreachable wrist poses [11–15].

The key design choice in SynManDex is to use human priors before contact, not at contact. A pre-grasp preserves approach direction, hand role, and coarse finger coordination while leaving the exact contact geometry to the robot embodiment. Human hand-object data is therefore useful not as direct robot supervision, but as a proposal distribution that places robot-native optimization in a functional basin.

We instantiate this idea as a three-stage synthetic data-generation pipeline for executable bimanual grasp-and-lift demonstrations. First, an object-conditioned diffusion model samples MANO pre-grasps from hand-object priors [16–19]. Second, these seeds are retargeted and refined by collision and force-closure objectives in the target robot configuration space. Third, arm-hand IK and dynamic rollouts admit only trajectories that survive approach, closure, squeeze, and lift. Across 312 objects from 25 classes, SynManDex achieves 86.4% force closure with 0.6 mm penetration and a 4.67/5 combined human-likeness audit score. Policies trained on the admitted trajectories reach 80.7% held-out simulated success and 25/30 successful zero-shot hardware trials on a three-object benchmark. Our experiments first isolate the seed-and-refine mechanism, then compare against retargeting, taxonomy, and optimization baselines under matched execution filters, and finally evaluate whether the admitted trajectories support policy learning and hardware transfer.

Contributions.

- We propose a human-to-robot grasp synthesis pipeline that uses generated MANO pre-grasps as pre-contact semantic seeds rather than executable labels.
- We develop a robot-native grounding stage that retargets these seeds and filters the resulting grasps through collision, force-closure, arm-hand IK, and dynamic lift checks.
- We build a validated bimanual demonstration dataset and policy-learning evaluation showing that the admitted trajectories improve simulated and hardware dexterous grasp execution.

Table 1 separates capability strength from evidence type, highlighting the missing intersection between human/task priors and robot-executable bimanual action data.

Method family	Phys. guid.	Sem. guid.	Task orient.	Cross embod.	Agentic compat.	Bimanual	H2R data
Physics-first grasp synthesis [3, 5, 6]	■■■■	–	■□□□	■■■□	■□□□	–	–
Semantic / task-oriented grasp synthesis [8, 10]	■■□□	■■■■	■■■□	■■□□	■□□□	–	■■□□
Bimanual robot grasp/data systems [20–22]	■■■■	■□□□	■■□□	■□□□	■■■□	■■■■	■□□□
Human-to-robot manipulation transfer [23]	■■■□	■■□□	■■■□	■■□□	■■■□	■■■■	■■■□
SynManDex (ours)	■■■■	■■■□	■■■□	■■■□	■■■□	■■■■	■■■□

Table 1. Capability stack from static dexterous grasp synthesis to human-guided, agent-compatible robot data. Rows summarize representative method families rather than exhaustive per-paper rankings. Each cell reports an ordinal 0–5 capability score: 5 is a core contribution with direct validation in the modality required by the column, 3 is explicit but partially validated support, 1 is incidental relevance, and – is outside scope. Color denotes evidence type, not score magnitude: green for robot-native physical grounding or execution, blue for semantic/task/human-intent guidance, violet for trajectory, policy, or downstream-agent compatibility, and amber for task relevance inferred from static grasp coverage.

2 Related Work

2.1 Human-Inspired Robotic Manipulation

Human hand-object data provides useful priors for hand pose, contact regions, and object affordances. HOI datasets are often parameterized with MANO [16]; GRAB [17], ContactPose [24], and DexYCB [25] encode complementary views of hand pose, object geometry, and contact. The question for robot learning is how this structure should enter a robot pipeline. Behavior cloning and retargeting systems can imitate human hand behavior [11–14], but direct imitation is sensitive to distribution shift and embodiment mismatch. Functional retargeting methods specify object- or task-centric objectives [26, 23], yet they still require task rewards and may produce local motions that do not preserve the intended contact when the robot morphology cannot realize the human pose.

Generative HOI models synthesize diverse digital human interactions [18, 27, 19]. They are attractive because they can produce object-conditioned task semantics without collecting new robot data, but their outputs remain MANO-space motions rather than robot-executable arm-hand trajectories. If those motions are retargeted directly, contacts can move, fingers can penetrate, and wrists can become unreachable. SynManDex places the human prior earlier in the pipeline: it samples pre-contact frames as semantic seeds and then lets robot-native optimization decide the physically valid contact state.

2.2 Dexterous Prehensile Manipulation

Dexterous grasp poses serve as keyframes for downstream object interactions, so they must satisfy both physical stability and semantic utility. Online reinforcement learning can discover dexterous behaviors [1, 2, 28], but contact rewards alone do not guarantee force closure or arm-hand reachability. Physics-aware grasp generators prioritize stability against gravity [3, 5, 6], and geometric retargeting improves correspondence across hands [29]; however, neither mechanism directly chooses object-function contacts. Recent bimanual systems and generative models scale data and improve pose diversity [22, 15, 30–33], but their utility for policy learning still depends on how samples are admitted as executable actions.

Vision-language and affordance-driven methods capture task intent via language instructions or open-world region matching [34, 9, 10]. These signals are useful for object-function reasoning, but they are usually too coarse to determine precise finger-level contact mechanics. Taxonomy-based approaches provide structured grasp templates [8], but still require embodiment-specific grounding. SynManDex separates these roles: the human model proposes a basin, and the robot optimizer tests contact, IK, and rollout feasibility. Table 1 summarizes this positioning, and Appendix G provides an expanded comparison.

3 Method

3.1 Problem Definition

We formalize bimanual dexterous grasp synthesis over left and right hands $l \in \{L, R\}$. Let the arm-hand configuration for each side be $\mathbf{q}_l = (\mathbf{T}_l, \boldsymbol{\theta}_l)$, where $\mathbf{T}_l$ is the wrist pose and $\boldsymbol{\theta}_l$ denotes the hand joints, yielding the bimanual configuration space $\mathcal{Q}_{\text{bi}} = \mathcal{Q}_L \times \mathcal{Q}_R$. Given an object mesh $\mathcal{M}$, SynManDex does not directly optimize a joint human-physics score. Instead, it uses the staged decomposition:

$$\begin{aligned} \mathbf{h}_0 &\sim p_\theta(\mathbf{h} \mid \mathcal{M}), \\ \mathbf{q}_{\text{init}} &= R_\psi(\mathbf{h}_0, \mathcal{M}), \\ \mathbf{q}^* &= \arg\min_{\mathbf{q} \in \mathcal{Q}_{\text{bi}}} w_c C_{\text{coll}}(\mathbf{q}, \mathcal{M}) + w_f \mathcal{L}_{\text{FC}}(\mathbf{q}, \mathcal{M}) + w_r \|\mathbf{q} - \mathbf{q}_{\text{init}}\|_2^2. \end{aligned} \tag{1}$$

Here $\mathbf{h}_0$ is a generated MANO pre-grasp, R_ψ retargets it to the robot embodiment, and the final optimization is performed in robot configuration space. Human priors therefore determine the initialization basin, while physical validity is evaluated by robot-native contact, collision, IK, and rollout checks [35, 36].

3.2 Human-to-Robot Optimization Framework

The pipeline has three stages. First, a diffusion model samples object-conditioned MANO pre-grasps that capture the intended approach and coarse finger coordination. Second, a geometric retargeting step maps each MANO pre-grasp to a robot seed while keeping the hand open enough for robot-native

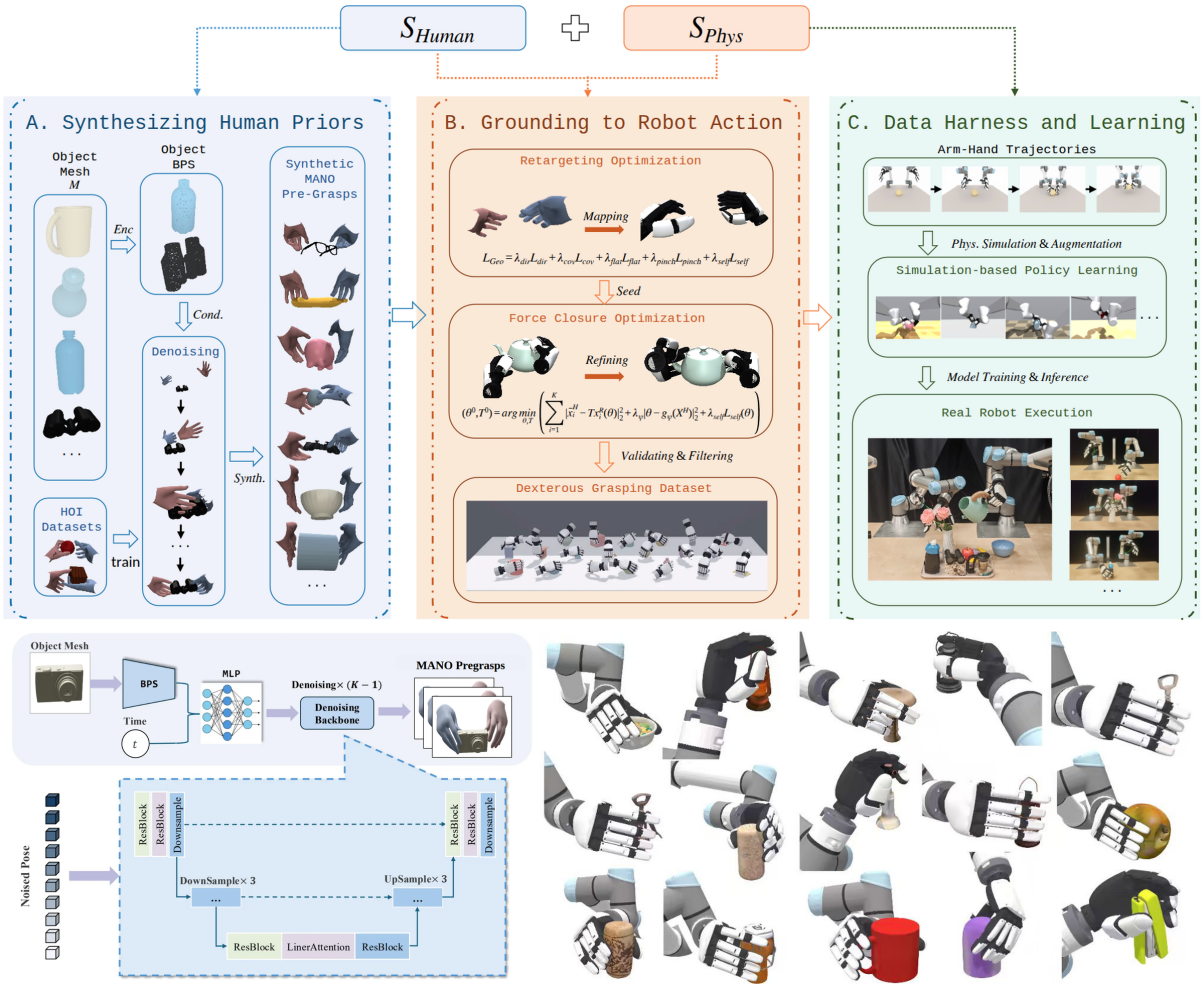

Figure 2. SynManDex pipeline. A diffusion model proposes object-conditioned MANO pre-grasps; geometric retargeting maps them to robot seeds; robot-native contact optimization and IK/rollout checks admit only executable grasp-and-lift trajectories for policy learning.

contact formation. Third, collision and force-closure optimization refines the seed on the target hand, after which arm-hand IK and rollout checks decide whether the sample is admitted as an executable demonstration. The complete pipeline is illustrated in Figure 2.

Human Prior Synthesis. To extract human priors from digital hand-object data, we introduce SynManDex-Human, an object-conditioned diffusion model tailored for hand-object interactions. We train on existing resources such as GRAB and ContactPose [16, 17, 24]. For temporal interaction sequences, we identify the first-contact frame by the minimum MANO-object distance and use a frame 0.2 s earlier as a pre-contact target. Static grasp records are used as hand-object pose priors rather than as temporal pre-contact sequences. This pre-contact supervision is intentional: it preserves approach direction and coarse coordination without forcing the robot to reproduce a human contact state that may be infeasible for its morphology. Appendix B.1 details the diffusion model architecture.

Mapping to Robot Pre-Grasps. Following denoising, the model yields a MANO seed $\mathbf{X}^H$ comprising $H = 21$ digital hand keypoints. Due to morphological and joint-space discrepancies between the human hand and a D_R-DoF target robotic embodiment (e.g., $D_R = 12$ for the XHand), direct joint-by-joint copying is infeasible. The retargeting stage is not meant to close the robot hand on the object. Its purpose is to place the robot wrist and fingers near the human-proposed approach basin while avoiding gross self-collision. Denoting the corresponding robotic keypoint parameterization as $\mathbf{x}_i^R(\boldsymbol{\theta})$, we employ a mapping g_ψ inspired by GeoRT [29] to preserve local motion geometry, spatial coverage, and self-collision margins. To maintain the functional semantics of a pre-grasp rather than a fully closed configuration, we relax the mapped pose toward an open-hand, taxonomy-compatible seed while aligning it in the object frame. The retargeting loss and objective are formulated as:

$$\mathcal{L}_{\text{GeoRT}} = \lambda_{\text{dir}}\mathcal{L}_{\text{dir}} + \lambda_{\text{cov}}\mathcal{L}_{\text{cov}} + \lambda_{\text{flat}}\mathcal{L}_{\text{flat}} + \lambda_{\text{pinch}}\mathcal{L}_{\text{pinch}} + \lambda_{\text{self}}\mathcal{L}_{\text{self}}, \tag{2}$$

$$(\boldsymbol{\theta}^0, \mathbf{T}^0) = \arg\min_{\boldsymbol{\theta},\mathbf{T}} \sum_i \|\bar{\mathbf{x}}_i^H - \mathbf{T}\mathbf{x}_i^R(\boldsymbol{\theta})\|_2^2 + \lambda_\psi \|\boldsymbol{\theta} - g_\psi(\mathbf{X}^H)\|_2^2 + \lambda_{\text{self}}\mathcal{L}_{\text{self}}(\boldsymbol{\theta}). \tag{3}$$

This optimization yields the structurally refined initial configuration for the subsequent force-closure stage.

Force-Closure Optimization. Initialized with $\mathbf{q}_{\text{init}} = (\mathbf{T}^0, \boldsymbol{\theta}^0)$, the pipeline executes a low-level optimization over contact kinematics and geometry. The human seed establishes a functional optimization basin, and the robot-native optimizer then reduces interpenetration, anchors feasible local contacts, and improves force closure. The optimized wrist-hand parameters are derived via:

$$\mathbf{q}^* = \arg\min_{\mathbf{q}} \; w_c C_{\text{coll}}(\mathbf{q}) + w_f \mathcal{L}_{\text{FC}}(\mathbf{q}) + w_r \|\mathbf{q} - \mathbf{q}_{\text{init}}\|^2, \tag{4}$$

$$Q_{\text{FC}}(\mathbf{q}) = \min_{\|\mathbf{w}\|=1} \max_{\mathbf{f}\in\mathcal{F}} \mathbf{w}^\top \mathbf{G}(\mathbf{q})\mathbf{f}. \tag{5}$$

where $\mathcal{L}_{\text{FC}} = -Q_{\text{FC}}$, and $\mathbf{G}$ maps friction-cone forces to the object wrench [35, 36]. Candidates are first filtered by penetration limits and force-closure thresholds. Because the force-closure metric is computed under a discretized friction-cone model, we treat it as an admission criterion rather than a guarantee of real-world stability. Surviving candidates then undergo dynamic floating-hand stability checks in Isaac Lab [37]; a grasp is retained only if it can support the object against gravity, adding a dynamic lift check beyond the static contact score.

Physical Validation Grounding. The preceding stages output isolated, floating wrist-hand grasp candidates. However, physical execution requires full kinematic reachability, collision-free approaches, and dynamic lifting capability. The default admitted sample is therefore a grasp-and-lift trajectory with approach, closure, squeeze, and lift phases. We employ cuRobo [38] to solve arm-hand inverse kinematics (IK) and simulate these trajectories in SAPIEN [39], adopting the UltraDexGrasp protocol [22]. The IK solver optimizes for an executable action sequence:

$$(\mathbf{a}_l^\star, \boldsymbol{\eta}_l^\star) = \arg\min_{\mathbf{a}_l,\boldsymbol{\eta}_l} \mathcal{L}_{SE(3)} + \lambda_{\text{kpt}}\mathcal{L}_{\text{kpt}} + \lambda_{\text{coll}}\mathcal{L}_{\text{coll}}^{\text{IK}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}^{\text{IK}}, \tag{6}$$

$$\mathcal{L}_{\text{kpt}} = \sum_i \rho\Big(\|\mathbf{T}_l^A(\mathbf{a}_l)\mathbf{x}_i^R(\boldsymbol{\theta}_l) - \mathbf{y}_{l,i}^*\|_2\Big), \tag{7}$$

$$y = \mathbf{1}\Big[\max_{t\geq t_{\text{lift}}}(z_t - z_0) > \tau_z\Big]. \tag{8}$$

Valid IK solutions are dynamically rolled out across approach, closure, squeeze, and lift phases. Only trajectories resulting in a successful vertical lift ($y = 1$) are admitted into the final imitation dataset. Extended manipulation tasks in Section 4.5 reuse these keyframes as initial possession states and add task-specific transitions.

3.3 Prehensile Manipulation Data Engine

Once a grasp successfully traverses the physical and kinematic filters, it serves as a primitive for downstream tasks. We use the admitted grasp-and-lift rollouts to train a closed-loop point-cloud policy and reuse selected validated keyframes as stable contact states for extended object-centric manipulation.

Demonstrations and Policy. For each admitted rollout, we store the object identifier, hand assignment, phase boundaries, optimized keyframe, validation metrics, and synchronized arm-hand states. These rollouts form the imitation dataset for a closed-loop, receding-horizon point-cloud policy operating on the union of scene geometry and rendered robot proprioceptive points:

$$\mathcal{D} = \{(\mathbf{P}_t, \mathbf{a}_{t:t+H-1})\}, \qquad \mathbf{P}_t = \mathbf{P}_t^{\text{scene}} \cup \mathbf{P}_t^{\text{robot}}, \tag{9}$$

$$p_\varphi(\mathbf{a}_{t:t+H-1} \mid \mathbf{P}_t) = \prod_{\tau=0}^{H-1} \text{TN}(\mathbf{a}_{t+\tau}; \boldsymbol{\mu}_{t+\tau}, \boldsymbol{\sigma}_{t+\tau}, \mathbf{a}_{\min}, \mathbf{a}_{\max}), \tag{10}$$

$$\mathcal{L}_{\text{policy}} = -\sum_{\tau=0}^{H-1} \log p_\varphi(\mathbf{a}_{t+\tau} \mid \mathbf{P}_t). \tag{11}$$

Visual features extracted via PointNet++ [40] are combined with action queries to predict bounded joint-trajectory chunks, ordered deterministically across the bimanual system (left-arm, left-hand, right-arm, right-hand). Rather than executing open-loop trajectories, the policy replans at each short horizon based on the most recent point-cloud observation. Full architectural details and hyperparameters are provided in Appendix J.

Manipulation. For extended object-centric manipulation, a validated grasp keyframe acts as the stable contact state for planning subsequent task motions. A task specification defines the object-relative transition, active and stabilizing hand assignments, and release conditions. We then use the established IK, collision, force-closure, and possession checks to admit transitions that maintain object control throughout the commanded motion. The application suite in Section 4.5 evaluates selected tasks–in-grasp reconfiguration, bimanual handovers, and pick-and-place–while separating keyframe validity from dynamic transition success.

4 Experiments

The experiments are organized as a staged test of the central claim: human priors are useful only when they are converted into robot-native, executable demonstrations. We therefore proceed from mechanism-level validation to executable simulation and hardware verification. Section 4.2 first isolates how the human seed and force-closure refinement interact; Section 4.3 then compares the resulting grasp generator against taxonomy, optimization, and retargeting baselines under the same execution filters. Section 4.4 asks whether the curated trajectories improve closed-loop policy learning. Sections 4.5 and 4.6 evaluate whether the same data supports manipulation beyond vertical lifting and transfers to hardware. Finally, Section 4.7 checks whether the seed-and-refine mechanism remains useful for a different dexterous hand. Detailed manifests, thresholds, and evaluator definitions are provided in Appendices H, H.1, I and J.

4.1 Setup

We instantiate this evaluation across a progressive embodiment stack: MANO space [16] $\rightarrow$ floating-base XHand $\rightarrow$ fully actuated 36-DOF bimanual UR5e-XHand system ($\mathcal{Q}_{\text{bi}}$). Each optimization query uses 240 parallel cuRobo seeds [38], followed by dynamic floating-hand stability checks in Isaac Lab [37]. Unless a table states otherwise, success requires the same admission chain: feasible contact, low collision/penetration, positive force closure, arm-hand IK, and a 10 cm lift or task-specific terminal-possession check. This convention makes the reported numbers stricter than static visual plausibility. We report generator-level metrics (penetration, force closure, IK/lift admission) separately from policy-level rollout success; the latter is used only in Sections 4.4 and 4.6. The 312-object, 25-class manifest is used for grasp-quality evaluation, while the 30 physical trials in Section 4.6 are a separate three-object hardware benchmark.

Dataset Coverage and Curation. Figure 3 provides a qualitative overview of the generated bimanual demonstrations before the section turns to controlled comparisons. The gallery is not used as evidence by itself; it shows the spatial and semantic coverage of the candidate pool, while admission into the policy dataset is controlled by the physical, kinematic, and dynamic criteria defined in Section 3.3 and Appendix H.1.

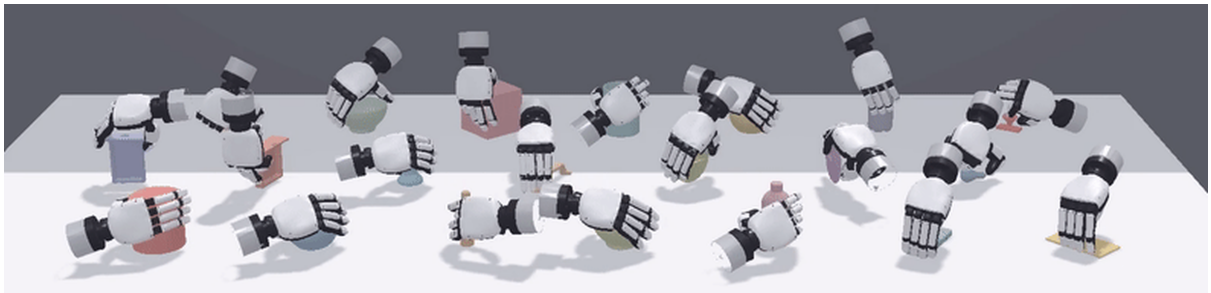

Figure 3. Qualitative coverage of generated bimanual demonstrations. A representative frame from the animated dataset gallery visualizes the semantic diversity of object-conditioned hand configurations produced by SynManDex. These examples show coverage, while admission into the policy training set requires the force-closure, penetration, IK, and dynamic lift checks defined in Appendix H.1.

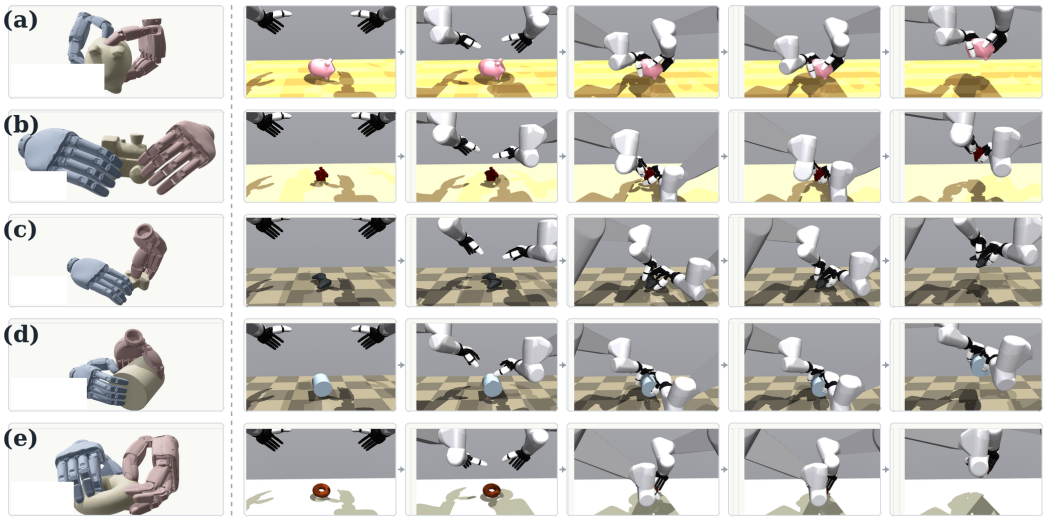

Figure 4. Trajectory grounding translates optimized keyframes into executable demonstrations. Rows (a)–(e) show piggy-bank, rose, duck, cylinder, and donut demonstrations. In each row, the left panel is the optimized goal pose, followed by the dynamic rollout sequence: start, approach, pre-grasp, grasp, and lift. Only trajectories satisfying the strict lift condition (Equation (8)) enter the imitation dataset.

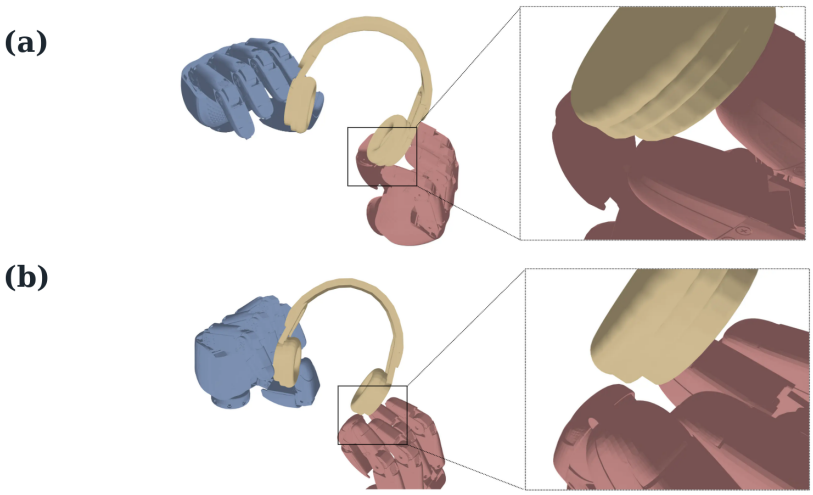

Figure 5. Force-closure refinement resolves contact-level failure after retargeting. Panel (a) shows the grasp after force-closure refinement, where local contact is grounded on the XHand geometry. Panel (b) shows the corresponding failure without this refinement, where the transferred pose remains visually plausible but misses the contact configuration needed for stable support.

4.2 Human Priors and Grasp Refinement

We begin with the mechanism that distinguishes SynManDex from direct retargeting or pure optimization: a human prior proposes the functional contact basin, and robot-native refinement decides whether that basin can be executed by the target embodiment. The qualitative progression in Figures 4 and 5 separates these two gates. Retarget-only preserves a plausible human-like hand shape but leaves contact unresolved; the full pipeline preserves the intended grasp role while reducing penetration and raising force closure in Table 2. Additional pre-grasp diagnostics are provided in Appendix C.

Figure 4 shows why this section evaluates trajectories rather than static poses: after contact refinement, every accepted keyframe must still survive approach, closure, and lift under the same arm-hand embodiment used by the policy. Figure 5 complements this rollout view by showing the local contact correction imposed by the force-closure objective in Equation (4).

Functional prior stress tests. Aggregate metrics do not reveal whether the human prior is preserving task intent, so we next place each representative failure beside the corresponding SynManDex recovery. Figure 6 tests functional bimanual priors for medium-size objects: fixed-size bimanual optimization

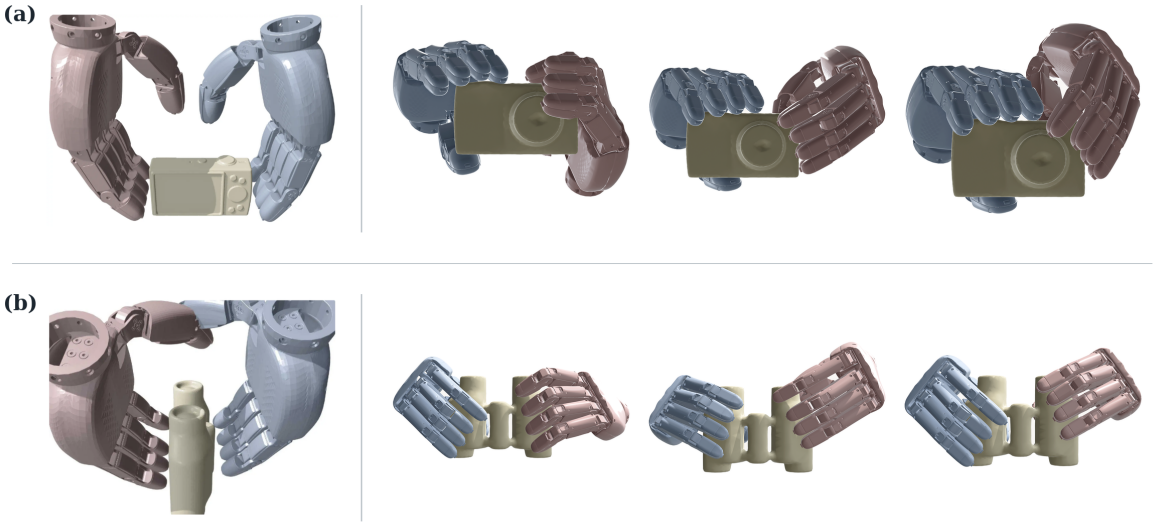

Figure 6. Functional bimanual consistency with failures placed beside successes. Rows (a) and (b) compare camera and binoculars grasps, respectively. In each row, the image left of the vertical separator is the other-method result; the images right of the separator are three SynManDex samples generated from the same human prior.

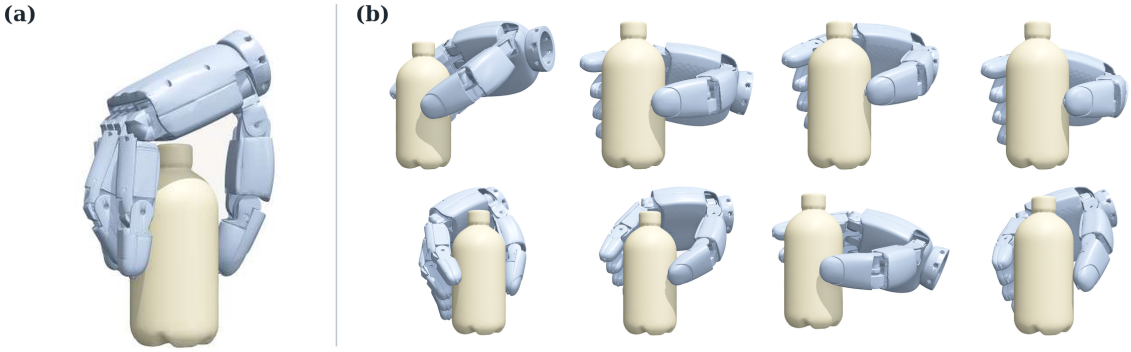

Figure 7. Direction stability for bottle side grasps. Panel (a) shows the BODex baseline, which produces a gripper-like unimanual wrap. Panel (b) shows eight SynManDex samples from a side-grasp human prior, preserving the intended side-oriented human-like grasp direction.

fails on camera and binoculars, whereas the same human prior yields multiple physically grounded SynManDex samples. Figure 7 isolates direction control in the unimanual setting, contrasting a BODex wrap with repeated side-grasp samples from SynManDex. Figure 8 then tests a fine-grained prior: a generic bimanual flute hold can be stable without preserving a task-motivated hand-object configuration, while SynManDex grounds four flute-holding release modalities from the same pose family. The appendix expands this case into a structured finger-release taxonomy (Appendix E).

Table 2. Grasp quality and human-likeness on the 312-object, 25-class grasp-quality manifest. Physical metrics evaluate optimized XHand configurations; G1 denotes the scaled Ferrari–Canny wrench-space margin under the contact model and Pen. reports penetration in mm. Human-likeness metrics report blinded VLM and human audits on a 1–5 scale plus plausibility-constrained diversity (PCD, Appendix H.2).

	Physical quality				Human-likeness / diversity			
Method	G1 ↑	Pen. (mm) ↓	Contact (%) ↑	FC (%) ↑	VLM-H (1–5) ↑	Human-H (1–5) ↑	Comb.-H (1–5) ↑	PCD ↑
SynManDex (full)	**7.2**	**0.6**	**89.2**	**86.4**	**4.72**	**4.59**	**4.67**	**0.41**
Optimization-only	4.6	1.1	71.6	79.1	2.86	2.74	2.81	0.11
Retarget-only	0.4	8.3	34.7	12.3	4.28	4.03	4.18	0.09

Compared with *Optimization-only*, SynManDex improves G1 stability by 56.5%, decreases penetration by 45.5%, and maintains a higher combined human-likeness audit score (4.67/5.0). The ablations iden-

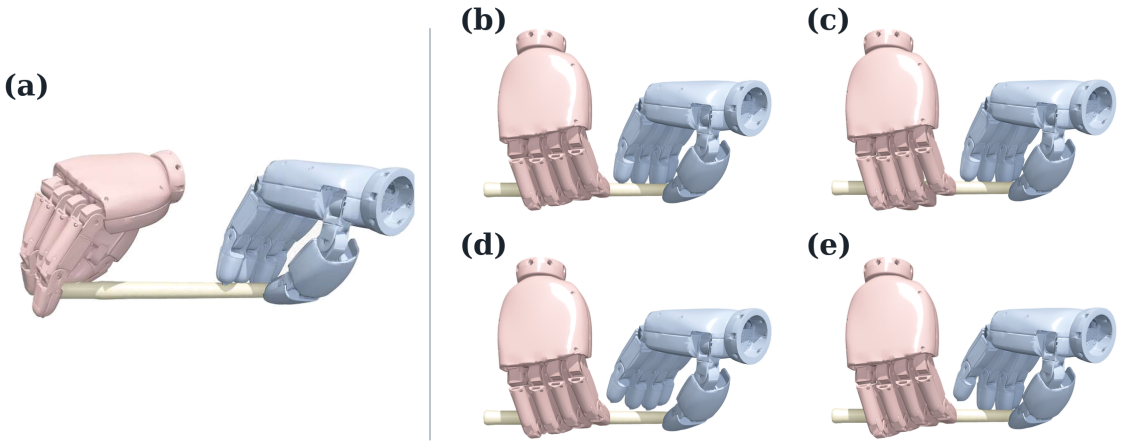

Figure 8. Flute-holding release modalities from a fine-grained human prior. Panel (a) shows a generic bimanual grasp baseline that can support the flute but does not preserve the task-motivated hand-object configuration. Panels (b)–(e) show SynManDex variants for all pressed, left-hand release, right-hand release, and cross-hand release modalities. The appendix organizes the larger set of flute variants into a structured finger-release taxonomy.

tify two complementary failure modes. *Retarget-only* isolates the embodiment gap: the visual prior is meaningful, but the transferred contact state is not physically grounded. *Optimization-only* isolates the semantic gap: the solver can find stable local contacts, but without a human seed it often converges to contact patterns that are unnatural or task-inappropriate. Among full-pipeline candidates, 82.3% satisfy IK reachability and 65.8% successfully execute the dynamic lift. Entropy, VLM, human-audit, and PCD protocols are detailed in Appendix H.2 and Table 10.

This section therefore validates the decoupling used throughout the paper. Human pre-grasps provide an affordance-consistent search basin, but they are not treated as robot grasps until geometry, contact normals, reachability, and lift execution are enforced. The remaining experiments reuse this same admission chain, so later policy, manipulation, and hardware results are measured on executable data rather than keyframe-level visual plausibility.

Diversity after physical filtering. High joint entropy alone can reward implausible or non-contact postures, so we audit diversity after the same physical filters have already been applied. Figure 9 shows that accepted SynManDex grasps remain diverse in hand-role assignment, approach direction, and object-relative support, complementing the PCD score in Tables 2 and 10.

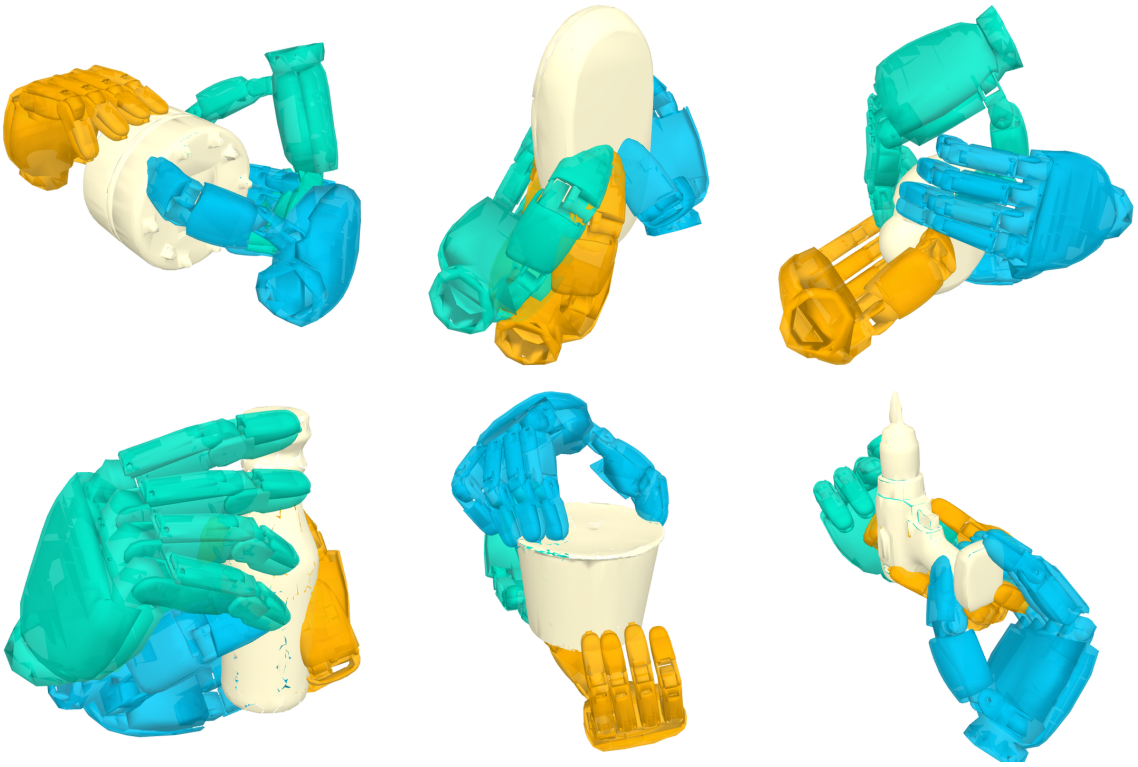

Figure 9. Qualitative diversity of accepted SynManDex grasps. The gallery shows six physically grounded bimanual grasps with different contact allocations, approach directions, and support roles. These examples complement the PCD score in Table 2 by showing that diversity is retained after strict physical and human-likeness filtering.

4.3 Benchmark and Taxonomy Comparisons

Having established the mechanism, we next ask whether it remains competitive under matched baselines rather than curated examples. The benchmark compares SynManDex with taxonomy, optimization, and retargeting paradigms under the same XHand contact, force-closure, IK, and lift filters. Dexonomy is formally adapted to the XHand embodiment [8]. Unimanual rows evaluate whether a method can realize a single functional grasp, while bimanual rows add the harder requirements of dual-hand coordination, collision-free reachability, and paired lift execution.

Table 3. Embodiment-aware grasp success. All methods use identical held-out object-task configurations; success is measured after force-closure filtering and arm-hand IK/lift execution.

Method	Prior / seed	FC (%) ↑	IK/lift (%) ↑	Task match (%) ↑
Unimanual grasping				
Dexonomy-XHand [8]	taxonomy	42.5	28.3	76.8
BODex [5]	optimization	71.4	39.6	45.0
dex-retargeting [11]	MANO pose	12.3	8.1	72.5
GeoRT [29]	geometric kpts.	38.9	24.7	68.2
SynManDex-Uni	**human prior + FC**	**86.4**	**72.9**	**81.7**
Bimanual grasping				
BODex-bimanual [5]	paired optimization	58.6	35.4	41.2
BODex-uni2bim [5]	uni → bim	49.7	26.8	43.9
dex-retargeting-bim [11]	bimanual MANO	10.8	6.5	74.1
GeoRT-bim [29]	bimanual kpts.	36.2	21.9	69.3
SynManDex	**human prior + FC**	**84.8**	**65.8**	**82.6**

Table 4. Comprehensive grasping benchmark. Static pose-only baselines are passed through the same SynManDex IK/rollout stack. IK/lift success is a generator-level admission metric and is distinct from the policy success reported in Section 4.4.

	Setup		Evaluation			
Method	Artifact	Bimanual	Pen. (mm) ↓	FC (%) ↑	Bench success (%) ↑	IK/lift (%) ↑
Dexonomy-XHand [8]	pose	no	4.7	42.5	36.8	28.3
DGN [3]	pose	no	3.4	54.8	46.2	33.9
BODex [5]	pose	no	1.4	74.6	63.5	45.7
UltraDexGrasp [22]	trajectory	yes	1.9	70.8	62.1	58.6
SynManDex	trajectory	yes	**0.6**	**86.4**	**78.9**	**65.8**

Tables 3 and 4 reinforce the mechanism-level diagnosis from Section 4.2. Taxonomy and direct MANO retargeting preserve task semantics better than random optimization, but they lose many samples at the embodiment-aware execution filters. Optimization-based baselines improve physical contact quality, yet they lag in task match and IK/lift admission when the search is not initialized inside a human-functional basin. SynManDex combines these two properties: human-prior task alignment and robot-native admission. Policy success is evaluated separately in Section 4.4.

4.4 Trajectory Data and Policy Learning

The preceding sections evaluate candidate generation; Table 5 evaluates whether those admitted trajectories are useful as policy data. Starting from the trajectory gate in Figure 4, we ablate the data source and the point-cloud policy interface while keeping the held-out object split, training budget, and terminal lift criterion fixed.

Table 5. Policy ablation study on the held-out simulated object split. Each row trains the same policy architecture while changing either the data source or the policy interface. Avg. L2 is the mean joint-space action error over executed chunks; Δ is relative to the full SynManDex policy.

Configuration	Success (%) ↑	Δ	Avg. L2 (rad) ↓
Full method (SynManDex policy)	**80.7**	—	**0.474**
Data source ablations			
No human prior (random-init optim.)	37.1	−43.6	0.622
No force closure (retarget-only)	22.9	−57.8	0.893
No Isaac Lab pre-validation	42.9	−37.8	0.561
Policy-interface ablations			
Scene-only point cloud	45.7	−35.0	0.539
MLP pooling without action queries	40.0	−40.7	0.601

The complete SynManDex policy reaches 80.7% success on held-out simulated objects. The data-source ablations show that performance is driven by structured demonstration quality, not simply by having kinematic paths: removing force-closure refinement drops success by 57.8 points, while removing the human prior drops success by 43.6 points. The policy-interface ablations then show that the controller also needs robot-aware observations and structured action queries for closed-loop correction. Protocol variants, including the open-loop diagnostic and GeoRT-IK ablation, are reported in Appendix H.1 and Tables 11 and 12.

4.5 Extended Manipulation Suites from Validated Grasp Keyframes

Figure 10. Selected in-grasp manipulation trials from validated keyframes. These simulated sequences test asynchronous coordination, where one hand maintains support while the other applies directional manipulation forces.

We next evaluate whether validated grasp keyframes can serve as initial possession states for selected object-centric transitions. These experiments are not intended to claim general-purpose manipulation; they evaluate whether contact states admitted by SynManDex remain useful under in-grasp reconfiguration, self-handover, and pick-and-place rollouts. Figure 10 evaluates in-grasp reconfiguration, Figure 11 shows object-specific bimanual prehensile grasps, and Figures 12 and 13 instantiate the release and transfer protocols summarized in Table 7.

Table 6. Simulated in-grasp manipulation evaluation. Metrics separate initial keyframe stability from contact-maintaining object reconfiguration.

	Initial state	Transition		
Method	Keyframe FC (%) ↑	Success (%) ↑	Slip (cm) ↓	Final check (%) ↑
Retarget-only	14.6	8.3	5.8	6.2
Static force-closure grasp	82.1	41.7	2.9	35.4
SynManDex	**87.5**	**70.8**	**1.1**	**66.7**

Table 6 separates static keyframe quality from dynamic transition success. Retarget-only grasps remain semantically plausible but often fail when force is applied because the contacts are not load-bearing. Static force-closure optimization stabilizes the initial state but can still slip during reconfiguration because the contact assignment is not informed by the later motion. SynManDex performs better because the stabilizing hand, active hand, and transition constraints are aligned before the trajectory enters the policy and manipulation suites.

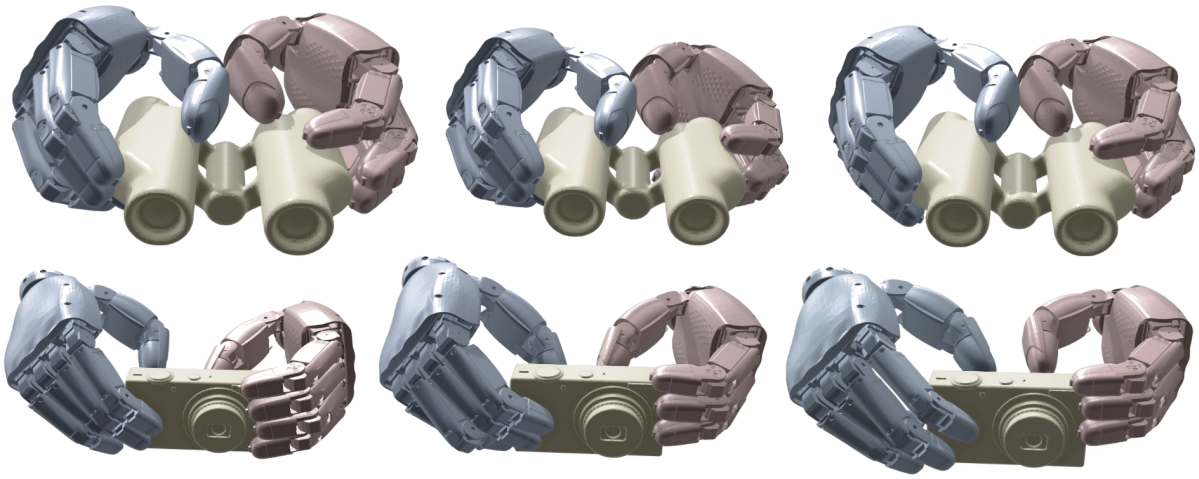

Figure 11. Object-specific bimanual prehensile grasps. Objects such as binoculars or cameras require dual-hand contact patterns conditioned on object geometry rather than generic power grasps.

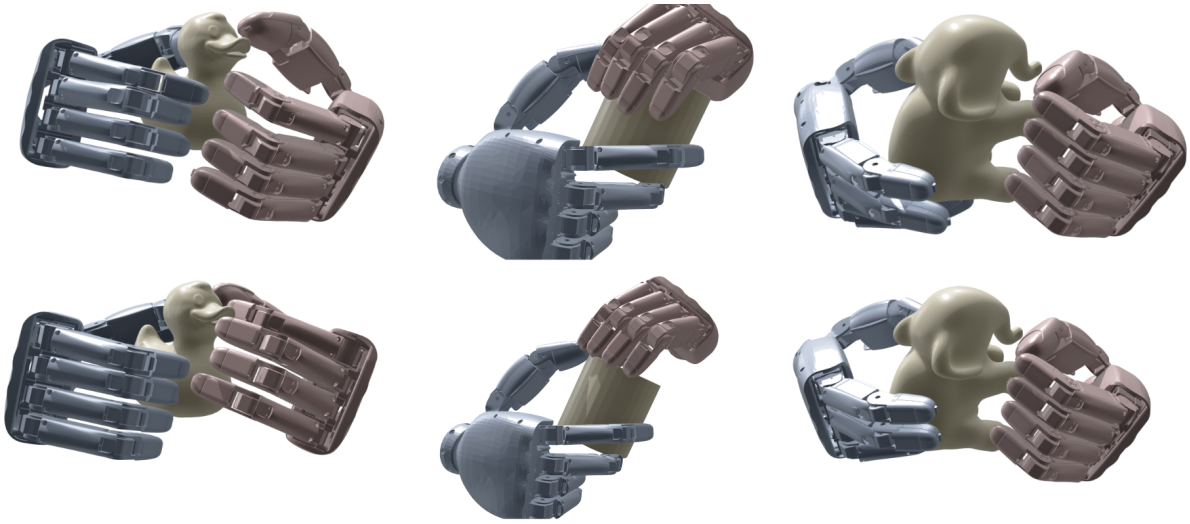

Figure 12. SynManDex creates handover-compatible bimanual keyframes. Diverse passing grasps establish compatible dual contacts on compact objects before one hand releases. These examples instantiate the bimanual handover rows in Table 7.

From Reconfiguration to Dynamic Transfer. Release transitions are a stricter test because one hand must disengage while the other assumes full possession and supports downstream placement. Figure 13 illustrates the pick-and-place suite, and Table 7 tracks the sequence from pre-release feasibility to terminal possession. We report Keyframe Validity for the pre-release grasp and End-to-End Success for the complete release, transfer, and placement rollout; trial definitions are specified in Appendix H.1.

Figure 13. Pick-and-place task illustration. The robot must establish contact, lift the object, transport it to the target region, and release or stabilize it without losing possession. This task instantiates the pick-and-place rows in Table 7.

Table 7. Simulated self-handover and pick-and-place evaluation. Keyframe validity measures pre-release feasibility; end-to-end success measures closed-loop possession through the terminal action.

Suite	Protocol	Trials	Keyframe valid (%) ↑	E2E success (%) ↑	Main failure slice
Bimanual handover	self handover, left→right release	24	87.5	70.8	release slip
	self handover, right→left release	24	83.3	66.7	asymmetric contact
Pick-and-place	tabletop pick → target zone	24	91.7	75.0	placement overshoot
	offhand support → placement	24	87.5	70.8	late offhand release
	cluttered pickup → bin drop	24	75.0	58.3	partial occlusion

The failure categories explain where the pipeline still loses executions. *Release slip* indicates insufficient support by the receiving hand before handoff; *asymmetric contact* reflects incompatible bilateral approach vectors; *placement overshoot* occurs after possession has already been established. The cluttered pickup condition is the hardest because partial occlusion degrades the point-cloud observation before the closed-loop policy can resolve a stable pre-grasp.

4.6 Sim-to-Real Hardware Validation

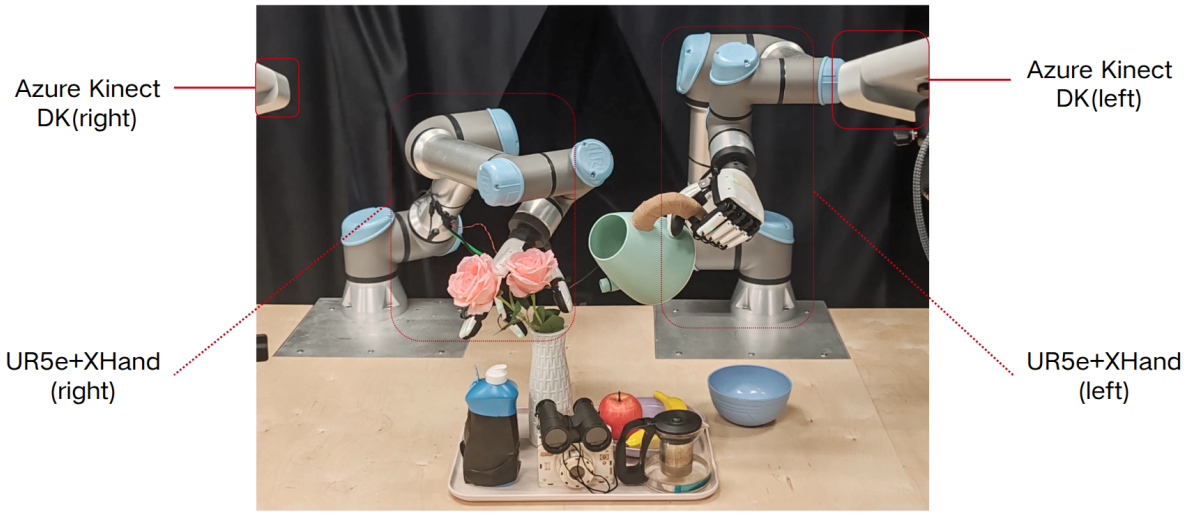

Figure 14. Real-system hardware validation platform. The bimanual UR5e+XHand platform is observed by two calibrated Azure Kinect DK cameras over a tabletop workspace.

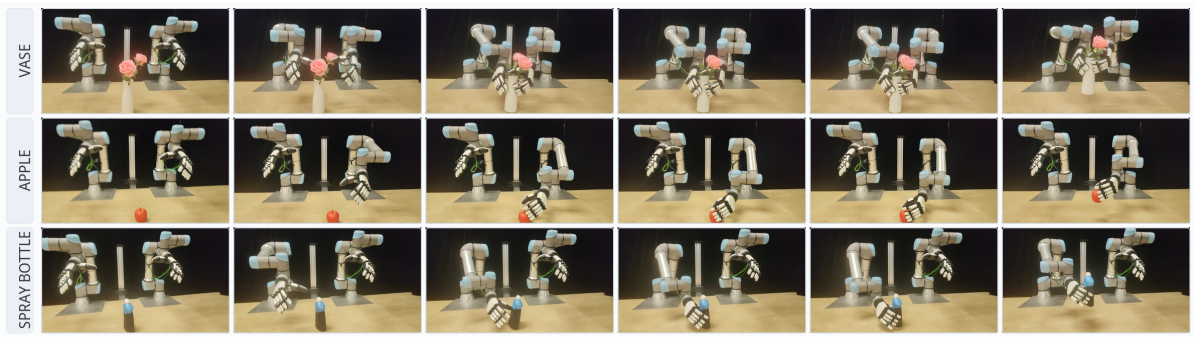

Figure 15. Zero-shot sim-to-real closed-loop hardware validation. Sequential execution strips show successful vase, apple, and spray-bottle examples from the three-object hardware benchmark, driven entirely by the simulation-trained point-cloud policy.

Together, Figures 14 to 16 move the same trajectory and policy stack from simulation to hardware. The quantitative hardware benchmark contains three everyday objects with ten trials per object; additional functional trials are shown qualitatively in Figure 16. The hardware validation mirrors the simulation interface: the bimanual UR5e-XHand platform receives fused point clouds and the policy replans short joint-trajectory chunks online (Appendix H.1). A physical trial is counted as successful only if the

system establishes contact, lifts or transports the object, and maintains possession at the commanded terminal state. On the three-object tabletop protocol, the SynManDex policy succeeds in 25/30 trials (8/10 vase, 8/10 apple, 9/10 spray bottle), while the recorded failure modes—rim slip, rotational shift at contact, and handle occlusion—match the simulation diagnostics.

Figure 16 adds three functional hardware trials that connect the manipulation studies in Section 4.5 to real execution. Camera lifting illustrates execution on an additional toy camera outside the three-object hardware table. Pick-handover-put aligns the real rollout with the simulated handover and pick-and-place protocols in Figures 12 and 13, verifying transfer, release, and terminal placement rather than only lift. Pouring illustrates a tilted functional terminal state, where possession must remain stable while the object expresses its intended use. Table 8 isolates the role of demonstration quality under the same hardware reset and evaluation protocol.

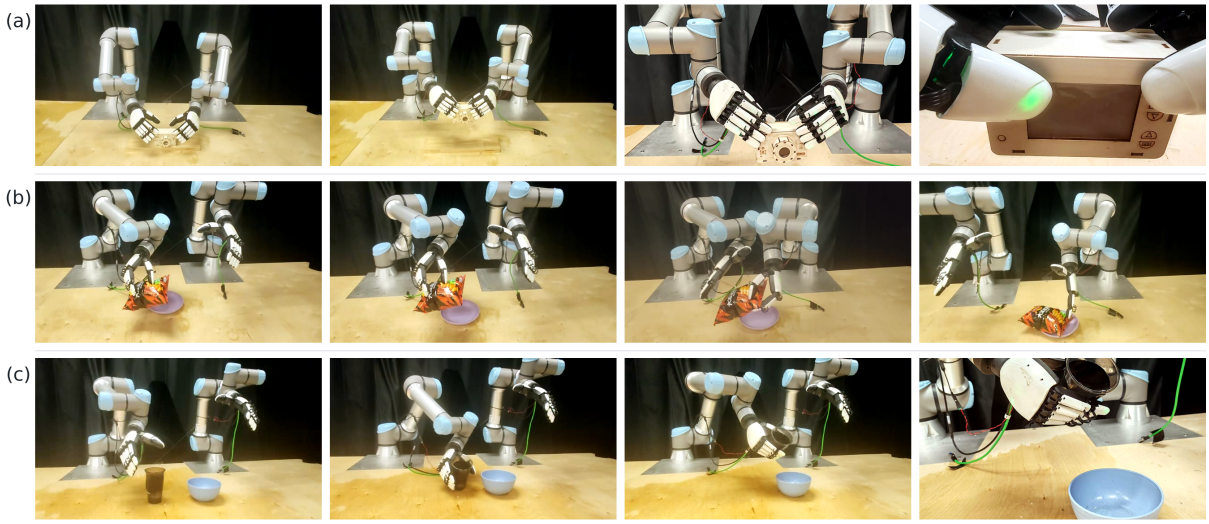

Figure 16. Extended real-world functional trials from video keyframes. Rows (a)–(c) show camera lifting, pick-handover-put, and pouring. In (a), the robot lifts an additional toy camera not included in the three-object hardware benchmark; the last two frames are third- and first-view after-lift close-ups. In (b), real execution follows the same staged structure as the simulated handover and pick-and-place protocols in Figures 12 and 13: pickup, transfer, handover, and placement. In (c), the robot maintains possession through a tilted terminal state for pouring.

Table 8. Hardware data-source comparison on the three-object benchmark. Each policy is trained with a different demonstration source and evaluated on the same 30 physical trials: 10 each for vase, apple, and spray bottle.

Data source	Trials	Success ↑	Lift valid ↑	Dominant failure mode
Retarget-only data-based policy	30	5/30 (16.7%)	6/30	unstable contact
Optimization-only data-based policy	30	11/30 (36.7%)	13/30	poor approach alignment
SynManDex data-based policy	30	**25/30 (83.3%)**	**26/30**	occlusion / slip

4.7 Cross-Embodiment Seeding Diagnostic

The final experiment returns to the mechanism itself: if human priors are useful because they initialize an affordance-consistent basin, they should also help a different dexterous hand before any XHand-specific policy learning is involved. We therefore replace the standard initialization of a BODex Shadow Hand optimization [5] with a MANO→Shadow geometric seed while keeping the object manifest, optimizer budget, collision parameters, and force-closure criteria fixed. This diagnostic does not establish morphology-agnostic policy transfer; it evaluates whether the pre-grasp seed distribution improves the contact-basin search of an existing Shadow-hand optimizer. Under this matched protocol, the semantic seed increases valid grasps from 96/384 to 142/384, raises force closure from 44.3% to 61.5%, and reduces penetration from 1.8 mm to 1.2 mm. Figure 17 and Table 9 summarize the evidence, with full protocol details in Appendix F.

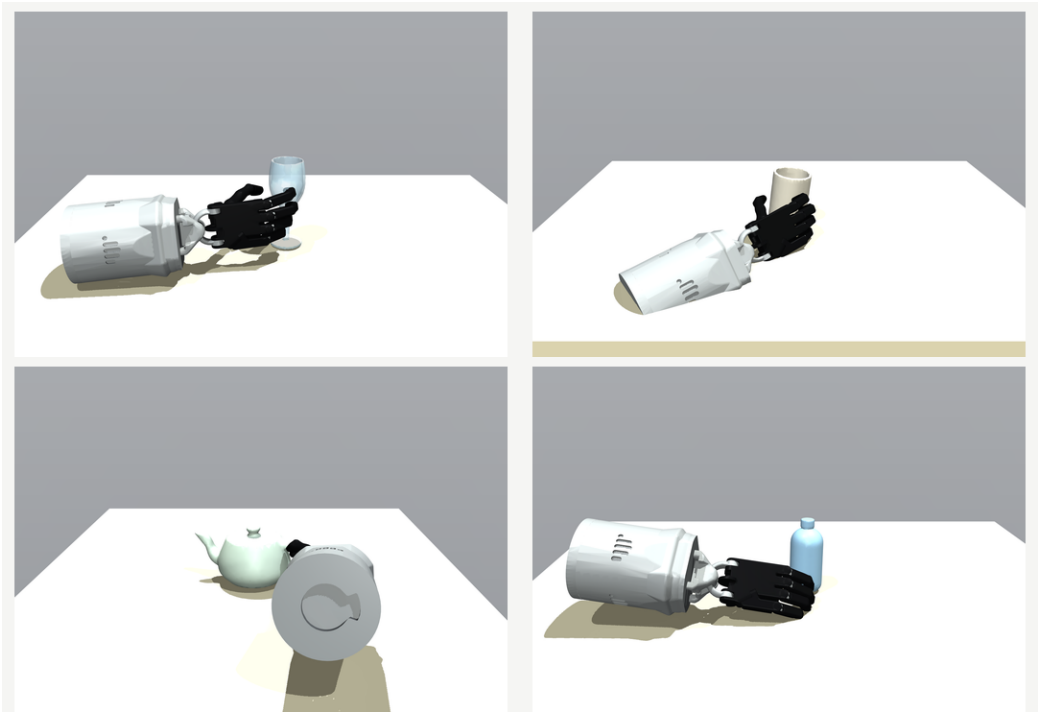

Figure 17. Shadow-hand seeding diagnostic. MANO→Shadow human-prior seeds place the Shadow wrist and fingers near object-relevant regions before BODex refinement, improving the candidate basin in this controlled setting.

Table 9. Cross-embodiment Shadow-hand diagnostic. Both rows use the same BODex solver, object manifest, and restart budget; only the seed changes. G1 is $Q_{\mathrm{FC}} \times 10^{-2}$ and penetration is in mm.

Method	Seed	Retarget err. ↓	Valid grasps ↑	FC (%) ↑	G1 ↑	Pen. ↓
BODex-Shadow random	native BODex	n/a	96/384	44.3	3.8	1.8
SynManDex-Shadow seeded	MANO→Shadow	7.6 mm	**142/384**	**61.5**	**5.1**	**1.2**

5 Limitations

Current SynManDex relies on visual geometry and proprioception. Force-closure optimization provides a useful static contact filter, but it does not model all dynamic effects, material variation, compliance, or tactile feedback. As a result, grasps can still slip during squeeze, release, or in-grasp reconfiguration. Incorporating tactile sensing and contact-state feedback is therefore an important next step.

The empirical scope is also limited. The hardware benchmark uses three objects with 30 trials, and the extended functional trials are qualitative. Human-likeness and diversity metrics rely partly on VLM and human audits, which are useful diagnostics but not substitutes for large-scale human preference studies or contact-distribution analysis. Finally, the Shadow-hand result is a controlled seeding diagnostic rather than a full cross-embodiment policy benchmark. These limitations suggest that future work should jointly model object geometry, embodiment reachability, tactile feedback, and execution-time contact robustness.

6 Conclusion

We presented SynManDex, a seed-and-refine pipeline that converts generated human pre-grasps into executable dexterous robot demonstrations. The central idea is to use human hand-object priors before contact, where they provide approach direction, hand role, and coarse finger coordination, and to leave final contact formation to robot-native collision, force-closure, IK, and rollout checks. Across grasp-quality benchmarks, policy ablations, and a three-object hardware evaluation, this separation improves both physical validity and human-like contact selection over retargeting-only or optimization-only alternatives. A Shadow-hand diagnostic further suggests that human-prior seeding can help contact-basin

discovery beyond the primary XHand embodiment. More broadly, SynManDex indicates a practical route for using human interaction priors without treating human contacts as directly executable robot labels.

References

[1] Aude Billard and Danica Kragic. Trends and challenges in robot manipulation. *Science*, 364(6446): eaat8414, 2019.

[2] OpenAI, Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020.

[3] Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang. DexGraspNet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. In *International Conference on Robotics and Automation (ICRA)*, pages 11359–11366. IEEE, 2023.

[4] Yinzhen Xu, Weikang Wan, Jialiang Zhang, Haoran Liu, Zikang Shan, Hao Shen, Ruicheng Wang, Haoran Geng, Yijia Weng, Jiayi Chen, Tengyu Liu, Li Yi, and He Wang. 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 (CVPR)*, pages 4737–4746, 2023.

[5] Jiayi Chen, Yubin Ke, and He Wang. Bodex: Scalable and efficient robotic dexterous grasp synthesis using bilevel optimization. In *International Conference on Robotics and Automation (ICRA)*, 2025.

[6] René Zurbrügg, Andrei Cramariuc, and Marco Hutter. GraspQP: Differentiable optimization of force closure for diverse and robust dexterous grasping. In *Conference on Robot Learning (CoRL)*, 2025.

[7] Thomas Feix, Javier Romero, Heinz-Bodo Schmiedmayer, Aaron M. Dollar, and Danica Kragic. The GRASP taxonomy of human grasp types. *IEEE Transactions on Human-Machine Systems*, 46(1): 66–77, 2016. doi: 10.1109/THMS.2015.2470657.

[8] Jiayi Chen, Yubin Ke, Lin Peng, and He Wang. Dexonomy: Synthesizing all dexterous grasp types in a grasp taxonomy. In *Robotics: Science and Systems (RSS)*, 2025.

[9] Yingbo Tang, Shuaike Zhang, Xiaoshuai Hao, Pengwei Wang, Jianlong Wu, Zhongyuan Wang, and Shanghang Zhang. Affordgrasp: In-context affordance reasoning for open-vocabulary task-oriented grasping in clutter. *arXiv preprint arXiv:2503.00778*, 2025.

[10] Yi-Lin Wei, Mu Lin, Yuhao Lin, Jian-Jian Jiang, Xiao-Ming Wu, Ling-An Zeng, and Wei-Shi Zheng. Afforddexgrasp: Open-set language-guided dexterous grasp with generalizable-instructive affordance. *arXiv preprint arXiv:2503.07360*, 2025.

[11] Yuzhe Qin, Yueh-Hua Wu, Shaowei Liu, Hanwen Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. DexMV: Imitation learning for dexterous manipulation from human videos. In *European Conference on Computer Vision (ECCV)*, pages 570–587. Springer, 2022.

[12] Ankur Handa, Karl Van Wyk, Wei Yang, Jacky Liang, Yu-Wei Chao, Qian Wan, Stan Birchfield, Nathan D. Ratliff, and Dieter Fox. DexPilot: Vision-based teleoperation of dexterous robotic hand-arm system. In *International Conference on Robotics and Automation (ICRA)*, pages 9164–9170, 2020.

[13] Priyanka Mandikal and Kristen Grauman. DexVIP: Learning dexterous grasping with human hand pose priors from video. In *Conference on Robot Learning (CoRL)*, pages 651–661, 2022.

[14] Shuqi Zhao, Xinghao Zhu, Yuxin Chen, Chenran Li, Xiang Zhang, Mingyu Ding, and Masayoshi Tomizuka. DexH2R: Task-oriented dexterous manipulation from human to robots. *arXiv preprint arXiv:2411.04428*, 2024.

[15] Juncheng Mu, Sizhe Yang, Yiming Bao, Hojin Bae, Tianming Wei, Linning Xu, Boyi Li, Huazhe Xu, and Jiangmiao Pang. DexImit: Learning bimanual dexterous manipulation from monocular human videos. *arXiv preprint arXiv:2602.10105*, 2026.

[16] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (ToG)*, 36(6):1–17, 2017.

[17] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, pages 581–600. Springer, 2020.

[18] Sammy Christen, Shreyas Hampali, Fadime Sener, Edoardo Remelli, Tomas Hodan, Eric Sauser, Shugao Ma, and Bugra Tekin. DiffH2O: Diffusion-based synthesis of hand-object interactions from textual descriptions. In *ACM SIGGRAPH Asia 2024 Conference Papers*, 2024. doi: 10.1145/3680528.3687563.

[19] Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. GOAL: Generating 4D whole-body motion for hand-object grasping. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13263–13273, 2022.

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

[21] Mu Lin, Yi-Lin Wei, Jiaxuan Chen, Yuhao Lin, Shuoyu Chen, Jiangran Lyu, Jiayi Chen, Yansong Tang, He Wang, and Wei-Shi Zheng. Bidexgrasp: Coordinated bimanual dexterous grasps across object geometries and sizes. *arXiv preprint arXiv:2604.06589*, 2026.

[22] Sizhe Yang, Yiman Xie, Zhixuan Liang, Yang Tian, Jia Zeng, Dahua Lin, and Jiangmiao Pang. UltraDexGrasp: Learning universal dexterous grasping for bimanual robots with synthetic data. *arXiv preprint arXiv:2603.05312*, 2026.

[23] Kailin Li, Puhao Li, Tengyu Liu, Yuyang Li, and Siyuan Huang. Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6991–7003, 2025.

[24] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *European Conference on Computer Vision (ECCV)*, pages 361–378. Springer, 2020.

[25] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9044–9053, 2021.

[26] Zhao Mandi, Yifan Hou, Dieter Fox, Yashraj Narang, Ajay Mandlekar, and Shuran Song. Dexmachina: Functional retargeting for bimanual dexterous manipulation. *arXiv preprint arXiv:2505.24853*, 2025.

[27] Hui Zhang, Sammy Christen, Zicong Fan, Otmar Hilliges, and Jie Song. GraspXL: Generating grasping motions for diverse objects at scale. In *European Conference on Computer Vision (ECCV)*, 2024.

[28] Ankur Handa, Arthur Allshire, Viktor Makoviychuk, Aleksei Petrenko, Ritvik Singh, Jingzhou Liu, Denys Makoviichuk, Karl Van Wyk, Alexander Zhurkevich, Balakumar Sundaralingam, Yashraj Narang, Jean-Francois Lafleche, Dieter Fox, and Gavriel State. DeXtreme: Transfer of agile in-hand manipulation from simulation to reality. In *International Conference on Robotics and Automation (ICRA)*, 2023.

[29] Zhao-Heng Yin, Changhao Wang, Luis Pineda, Krishna Bodduluri, Tingfan Wu, Pieter Abbeel, and Mustafa Mukadam. Geometric retargeting: A principled, ultrafast neural hand retargeting algorithm. *arXiv preprint arXiv:2503.07541*, 2025.

[30] Puhao Li, Tengyu Liu, Yuyang Li, Yiran Geng, Yixin Zhu, Yaodong Yang, and Siyuan Huang. GenDexGrasp: Generalizable dexterous grasping. In *International Conference on Robotics and Automation (ICRA)*, pages 8068–8074, 2023.

[31] Jialiang Zhang, Haoran Liu, Danshi Li, Xinqiang Yu, Haoran Geng, Yufei Ding, Jiayi Chen, and He Wang. DexGraspNet 2.0: Learning generative dexterous grasping in large-scale synthetic cluttered scenes. In *Conference on Robot Learning (CoRL)*, 2024.

[32] Guo-Hao Xu, Yi-Lin Wei, Dian Zheng, Xiao-Ming Wu, and Wei-Shi Zheng. Dexterous grasp transformer. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17933–17942, 2024.

[33] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *International Conference on Computer Vision (ICCV)*, pages 11107–11116, 2021.

[34] Quanzhou Li, Zhonghua Wu, Jingbo Wang, Chen Change Loy, and Bo Dai. Dhagrasp: Synthesizing affordance-aware dual-hand grasps with text instructions. *arXiv preprint arXiv:2509.22175*, 2025.

[35] Jiawei Li, Hang Liu, and He-Gao Cai. On computing three-finger force-closure grasps of 2-d and 3-d objects. *IEEE Transactions on Robotics and Automation*, 19(1):155–161, 2003. doi: 10.1109/TRA.2002.806774.

[36] Richard M Murray, Zexiang Li, and S Shankar Sastry. *A mathematical introduction to robotic manipulation*. CRC press, 1994.

[37] Mayank Mittal, Pascal Roth, James Tigue, Antoine Richard, Octi Zhang, Peter Du, Antonio Serrano-Munoz, Xinjie Yao, René Zurbrügg, Nikita Rudin, et al. Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning. *arXiv preprint arXiv:2511.04831*, 2025.

[38] Balakumar Sundaralingam, Siva Kumar Sastry Hari, Adam Fishman, Caelan Garrett, Karl Van Wyk, Valts Blukis, Alexander Millane, Helen Oleynikova, Ankur Handa, Fabio Ramos, Nathan Ratliff, and Dieter Fox. cuRobo: Parallelized collision-free minimum-jerk robot motion generation. *arXiv preprint arXiv:2310.17274*, 2023.

[39] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. SAPIEN: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11097–11107, 2020.

[40] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, 2017.

[41] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851, 2020.

[42] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. In *International Conference on Learning Representations (ICLR)*, 2023.

[43] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18000–18010, 2023.

[44] Yifan Han, Zhongxi Chen, Yuxuan Zhao, Congsheng Xu, Yanming Shao, Yichuan Peng, Yao Mu, and Wenzhao Lian. DexHiL: A human-in-the-loop framework for vision-language-action model post-training in dexterous manipulation. *arXiv preprint arXiv:2603.09121*, 2026.

[45] Sungjae Park, Seungho Lee, Mingi Choi, Jiye Lee, Jeonghwan Kim, Jisoo Kim, and Hanbyul Joo. Learning to transfer human hand skills for robot manipulations. *arXiv preprint arXiv:2501.04169*, 2025.

[46] Tao Chen, Jie Xu, and Pulkit Agrawal. A system for general in-hand object re-orientation. In *Conference on Robot Learning (CoRL)*, pages 297–307, 2022.

[47] Zhao-Heng Yin, Binghao Huang, Yuzhe Qin, Qifeng Chen, and Xiaolong Wang. Rotating without seeing: Towards in-hand dexterity through touch. In *Robotics: Science and Systems (RSS)*, 2023.

[48] Yuanpei Chen, Tianhao Wu, Shengjie Wang, Xidong Feng, Jiechuan Jiang, Zongqing Lu, Stephen McAleer, Hao Dong, Song-Chun Zhu, and Yaodong Yang. Towards human-level bimanual dexterous manipulation with reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2022.

[49] Chen Bao, Helin Xu, Yuzhe Qin, and Xiaolong Wang. Dexart: Benchmarking generalizable dexterous manipulation with articulated objects. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21190–21200, 2023.

[50] Jun Wang, Yuzhe Qin, Kaiming Kuang, Yigit Korkmaz, Akhilan Gurumoorthy, Hao Su, and Xiaolong Wang. CyberDemo: Augmenting simulated human demonstration for real-world dexterous manipulation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[51] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. In *Proceedings of Robotics: Science and Systems (RSS)*, 2018. doi: 10.15607/RSS.2018.XIV.049.

[52] Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *Conference on Robot Learning (CoRL)*, 2021.

[53] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems (RSS)*, 2023.

[54] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin C. M. Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems (RSS)*, 2023.

[55] Wenbo Zhang, Tianrun Hu, Yanyuan Qiao, Hanbo Zhang, Yuchu Qin, Yang Li, Jiajun Liu, Tao Kong, Lingqiao Liu, and Xiao Ma. Chain-of-action: Trajectory autoregressive modeling for robotic manipulation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[56] Xueyi Liu, Jianibieke Adalibieke, Qianwei Han, Yuzhe Qin, and Li Yi. Dextrack: Towards generalizable neural tracking control for dexterous manipulation from human references. In *International Conference on Learning Representations (ICLR)*, 2025.

[57] Bohan Zhou, Haoqi Yuan, Yuhui Fu, and Zongqing Lu. Learning diverse bimanual dexterous manipulation skills from human demonstrations. In *International Conference on Learning Representations (ICLR)*, 2025.

[58] Weikang Wan, Haoran Geng, Yun Liu, Zikang Shan, Yaodong Yang, Li Yi, and He Wang. 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 (ICCV)*, pages 3891–3902, 2023.

[59] Arsalan Mousavian, Clemens Eppner, and Dieter Fox. 6-dof graspnet: Variational grasp generation for object manipulation. In *International Conference on Computer Vision (ICCV)*, pages 2901–2910, 2019.

[60] Enric Corona, Albert Pumarola, Guillem Alenyà, Francesc Moreno-Noguer, and Gregory Rogez. GANHand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5031–5041, 2020.

[61] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J. Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *International Conference on 3D Vision (3DV)*, pages 333–344, 2020.

[62] Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. ContactOpt: Optimizing contact to improve grasps. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1471–1481, 2021.

[63] Haofei Lu, Yifei Dong, Zehang Weng, Florian T. Pokorny, Jens Lundell, and Danica Kragic. Grasping a handful: Sequential multi-object dexterous grasp generation. *IEEE Robotics and Automation Letters*, 10(11):11880–11887, 2025. doi: 10.1109/LRA.2025.3614051.

[64] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[65] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

Algorithm 1 SynManDex as proposal, refinement, and executable filtering.

Require: Object mesh $\mathcal{M}$, number of human-prior samples N, lift threshold τ_z

Sample MANO pre-grasps $\{\mathbf{h}_0^i\}_{i=1}^N \sim p_\theta(\mathbf{h}_0 \mid \mathcal{M})$
for $i = 1, \ldots, N$ **do**
  Retarget $\mathbf{h}_0^i$ to an XHand seed $\mathbf{q}_{\text{init}}^i$ with GeoRT-calibrated geometry (Equations (2) and (3))
  Refine $\mathbf{q}_{\text{init}}^i$ into $\mathbf{q}^{i*}$ by minimizing collision, force-closure, and seed-regularization losses (Equation (4))
  Reject $\mathbf{q}^{i*}$ if floating-hand stability, penetration, or force-closure checks fail
  Solve GeoRT-guided arm-hand IK on the same UR5e+XHand model used in execution (Equation (6))
  Execute approach–close–squeeze–lift; retain the trajectory if $\max_{t \geq t_{\text{lift}}}(z_t - z_0) > \tau_z$
end for
return IK-validated demonstrations $\mathcal{D}$ with source provenance and failure labels

This supplementary material records the details behind the main-paper claims. It contains the full problem formulation and algorithm (Appendix A), method details omitted from the main text (Appendix B), qualitative diagnostics for the pre-grasp pipeline, unimanual grasps, and flute-holding release priors (Appendices C to E), and a fixed-manifest Shadow-hand study that evaluates whether MANO priors remain useful under a non-XHand embodiment (Appendix F). It also gives the expanded related work, experiment protocols, domain randomization, policy details, and limitations.

A Full Problem Definition and Pipeline Algorithm

Let $\mathbf{q}_l = (\mathbf{T}_l, \boldsymbol{\theta}_l)$ denote the configuration of manipulator $l \in \{L, R\}$, where $\mathbf{T}_l \in SE(3)$ is the wrist pose and $\boldsymbol{\theta}_l \in \mathbb{R}^{n_{\text{hand}}}$ the hand joint configuration. Given an object mesh $\mathcal{M}$, SynManDex formulates human-like dexterous grasping over the bimanual space $\mathcal{Q}_{\text{bi}} = \mathcal{Q}_L \times \mathcal{Q}_R$ as a staged proposal-and-filtering problem rather than a single joint solver:

$$\begin{aligned} \mathbf{h}_0 &\sim p_\theta(\mathbf{h} \mid \mathcal{M}), \\ \mathbf{q}_{\text{init}} &= R_\psi(\mathbf{h}_0, \mathcal{M}), \\ \mathbf{q}^* &= \arg\min_{\mathbf{q} \in \mathcal{Q}_{\text{bi}}} w_c C_{\text{coll}}(\mathbf{q}, \mathcal{M}) + w_f \mathcal{L}_{\text{FC}}(\mathbf{q}, \mathcal{M}) + w_r \|\mathbf{q} - \mathbf{q}_{\text{init}}\|_2^2. \end{aligned} \tag{12}$$

Here $\mathbf{h}_0$ is a generated MANO pre-grasp, R_ψ maps it to a robot seed, and the final optimization is performed in robot configuration space. Human grasp priors are therefore used as initialization rather than as executable labels. Compared with random initialization, a human-prior seed starts in a functional region of $\mathcal{Q}_{\text{bi}}$, after which contact, collision, force-closure, IK, and rollout checks decide whether a sample is admitted. SynManDex operationalizes this insight in three stages (Figure 2): synthesizing human pre-grasps (Section 3.2), retargeting and force-closure optimization (Section 3.2), and trajectory generation with policy training (Section 3.2).

B Full Method Details

B.1 Human Pre-Grasp Diffusion

At each diffusion step, SynManDex-Human estimates the distribution of a less-noisy MANO sample from the current noisy sample and object mesh:

$$p_\theta(\mathbf{h}_{t-1} \mid \mathbf{h}_t, \mathcal{M}) = \mathcal{N}(\mathbf{h}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{h}_t, t, \mathcal{M}), \boldsymbol{\Sigma}_\theta(\mathbf{h}_t, t, \mathcal{M})). \tag{13}$$

Using the standard DDPM reparameterization [41], the network predicts noise:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t, \mathbf{h}_0, \boldsymbol{\epsilon}} \left[\left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{h}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}, t, \mathcal{M}) \right\|_2^2 \right]. \tag{14}$$

The model generates a single pre-contact frame. Downstream retargeting, contact optimization, and trajectory generation are responsible for robot feasibility.

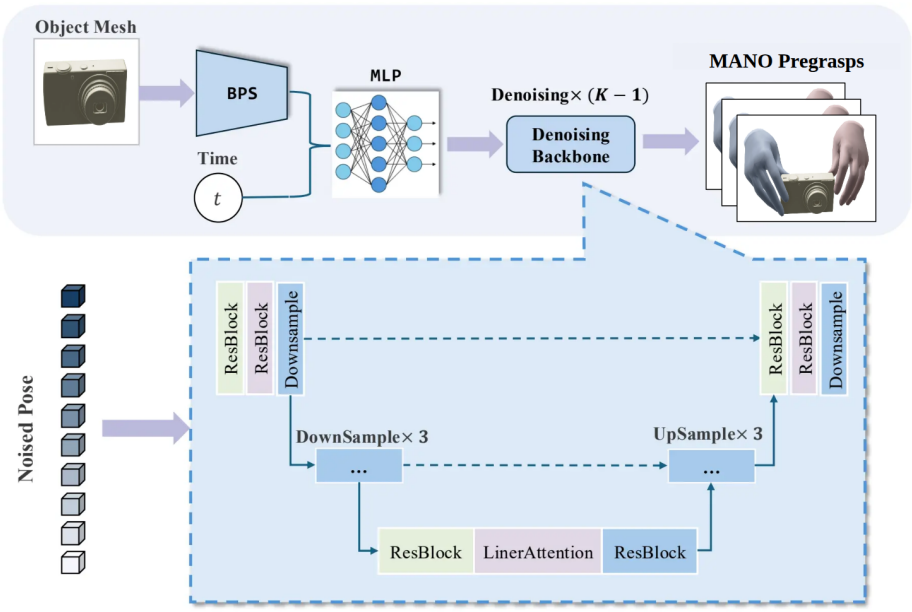

Figure 18. Architecture of SynManDex-Human. The object mesh is encoded via Basis Point Set (BPS) representation and projected through an MLP, which, together with the diffusion timestep t, conditions a U-Net denoising backbone. The backbone iteratively refines a noised pose into MANO pre-grasp parameters through $K-1$ denoising steps.

B.2 Retargeting Objective Terms

The compact retargeting objective in Equation (2) expands into five geometric criteria. Motion preservation aligns source and target keypoint displacement directions:

$$\mathcal{L}_{\text{dir}} = \sum_{i=1}^{K} \mathbb{E}_{\mathbf{X}^H, \delta_i} \left[1 - \frac{\Delta \mathbf{x}_i^R(\mathbf{X}^H, \delta_i)^\top \delta_i}{\|\Delta \mathbf{x}_i^R(\mathbf{X}^H, \delta_i)\|_2 \|\delta_i\|_2 + \epsilon} \right], \tag{15}$$

where $\Delta \mathbf{x}_i^R(\mathbf{X}^H, \delta_i) = \mathbf{x}_i^R(g_\psi(\mathbf{X}^H + \delta_i)) - \mathbf{x}_i^R(g_\psi(\mathbf{X}^H))$. Coverage uses a Chamfer proxy against a point-cloud approximation $\mathcal{P}_i^R$ of the XHand reachable keypoint space:

$$\mathcal{L}_{\text{cov}} = \sum_{i=1}^{K} \text{CD}\left(\{\mathbf{x}_i^R(g_\psi(\mathbf{X}_m^H))\}_{m=1}^{B}, \mathcal{P}_i^R \right). \tag{16}$$

Flatness penalizes nonuniform finite-difference response:

$$\mathcal{L}_{\text{flat}} = \sum_{i=1}^{K} \mathbb{E}_{\mathbf{X}^H, \delta_i} \left[\left\| \mathbf{x}_i^R(g_\psi(\mathbf{X}^H + \delta_i)) - 2\mathbf{x}_i^R(g_\psi(\mathbf{X}^H)) + \mathbf{x}_i^R(g_\psi(\mathbf{X}^H - \delta_i)) \right\|_2^2 \right]. \tag{17}$$

Pinch preservation keeps robot fingertips close when corresponding MANO fingertips are close:

$$\mathcal{L}_{\text{pinch}} = \sum_{(i,j) \in \mathcal{P}_{\text{pinch}}} \mathbf{1}\left[\|\mathbf{x}_i^H - \mathbf{x}_j^H\|_2 < \tau_H \right] \max\left(0, \|\mathbf{x}_i^R(g_\psi(\mathbf{X}^H)) - \mathbf{x}_j^R(g_\psi(\mathbf{X}^H))\|_2 - \tau_R \right)^2. \tag{18}$$

The self-collision term uses the same hand collision model used later for simulation filtering.

B.3 Force Closure and IK Details

For K contact points at positions $\{\mathbf{p}_k\}_{k=1}^K$ relative to the object center of mass, the grasp map is

$$\mathbf{G} = \begin{bmatrix} \mathbf{I}_3 & \cdots & \mathbf{I}_3 \\ [\mathbf{p}_1]_\times & \cdots & [\mathbf{p}_K]_\times \end{bmatrix}. \tag{19}$$

$Q_{\text{FC}} > 0$ certifies force closure under the discretized Coulomb friction cone. We approximate friction cones as 8-sided polyhedra and solve candidate refinements in parallel on GPU.

For GeoRT-IK, the wrist residual is

$$\mathcal{L}_{SE(3)} = \|\mathbf{p}_l^A(\mathbf{a}_l) - \mathbf{p}_l^*\|_2^2 + \lambda_R \|\operatorname{Log}(\mathbf{R}_l^* \mathbf{R}_l^A(\mathbf{a}_l)^\top)^\vee\|_2^2, \tag{20}$$

and pinch preservation maintains the fingertip pair distances from the optimized grasp:

$$\mathcal{L}_{\text{pinch}}^{\text{IK}} = \sum_{(i,j)\in\mathcal{P}_{\text{pinch}}^*} \left(\left\| \mathbf{T}_l^A(\mathbf{a}_l)\mathbf{x}_i^R(\boldsymbol{\theta}_l) - \mathbf{T}_l^A(\mathbf{a}_l)\mathbf{x}_j^R(\boldsymbol{\theta}_l) \right\|_2 - \|\mathbf{y}_{l,i}^* - \mathbf{y}_{l,j}^*\|_2 \right)^2. \tag{21}$$

Solutions are rejected if wrist residual, keypoint residual, collision margin, or joint-limit slack exceed thresholds.

C Qualitative Pre-Grasp Progression

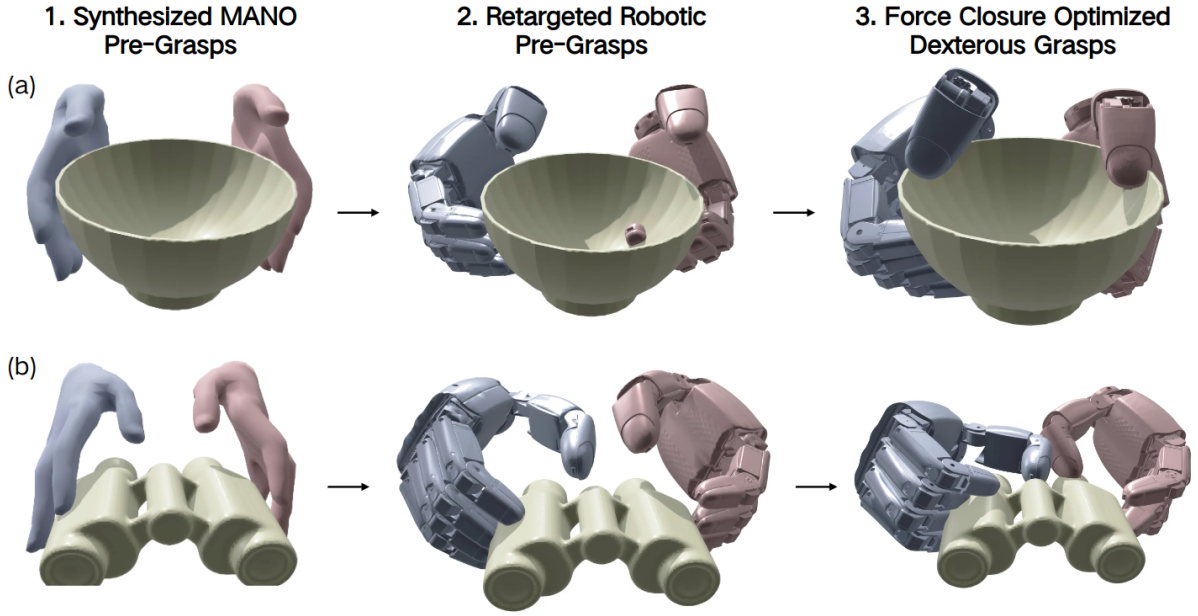

Figure 19. Pipeline progression for (a) a bowl and (b) binoculars. Left: SynManDex-Human synthesizes MANO pre-grasps that capture human grasp intuition: ergonomic approach directions and natural finger coordination. Center: GeoRT-calibrated retargeting (Equations (2) and (3)) preserves fingertip placement and pinch geometry across the 45→12 DoF gap, with residual object interpenetration and wrist offset. Right: SynManDex-Optimization (Equation (4)) resolves both, producing physically stable, prior-aligned grasps.

Figure 19 shows the qualitative mechanism behind the main refinement study. The MANO samples encode human-like approach and finger coordination; retargeting preserves much of that structure while leaving physical artifacts; force-closure refinement removes penetration while staying near the human-prior basin. This visual diagnostic complements Figure 5 and Table 2: retarget-only grasps have 8.3 mm penetration and 12.3% force-closure rate, while the full pipeline reaches 0.6 mm and 86.4%.

D Unimanual Grasp Generation

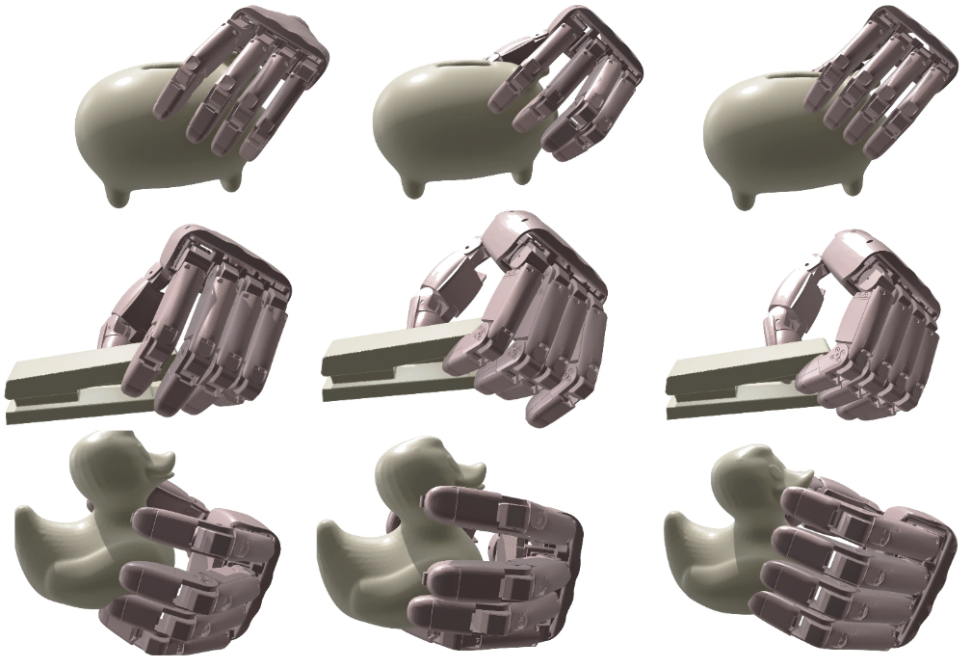

Figure 20. Unimanual grasps generated by SynManDex. Each row shows three diverse single-hand grasps on different objects: a piggy bank (top), a stapler (middle), and a rubber duck (bottom). Using only the single-hand subset of the human prior, SynManDex produces stable and natural unimanual grasps suitable for single-arm manipulation and object placement.

Although the main paper focuses on bimanual grasping, Figure 20 shows that the same human-prior-to-physics mechanism also works in the single-hand regime. The examples cover rounded, elongated, and irregular objects, producing palmar wrap, precision-power, and stabilizing grasps after force-closure refinement. This result is kept as an appendix diagnostic: it supports the generality of the representation, while the main empirical claim remains bimanual dexterous grasp generation.

E Finger-Release Taxonomy for Flute-Holding Grasp Poses

The main paper uses Figure 8 to compare a generic bimanual flute hold against four SynManDex flute-holding modalities. To better demonstrate the structure of the generated poses, we organize the larger flute set with a finger-release taxonomy. Starting from a canonical two-hand flute grasp, each pose is labeled by the subset of fingers released from the instrument, such as $L{:}\{I, M, R\}, R{:}\{\}$ for releasing the left index, middle, and ring fingers. This organization reflects a natural property of flute interaction: the hand must maintain instrument support while individual fingers change contact with the keys. We use this taxonomy as an interpretable visualization of grasp-pose diversity, not as a model of musical performance.

The release labels are generated by parsing the released-finger tags associated with each pose. For example, releasing the left index, middle, and ring fingers maps to the compact taxonomy key $L{:}\{I, M, R\}, R{:}\{\}$. We group poses first by release count, then by whether the released fingers belong to the left hand, right hand, or both hands. This produces a compact tree representation for 256 release combinations.

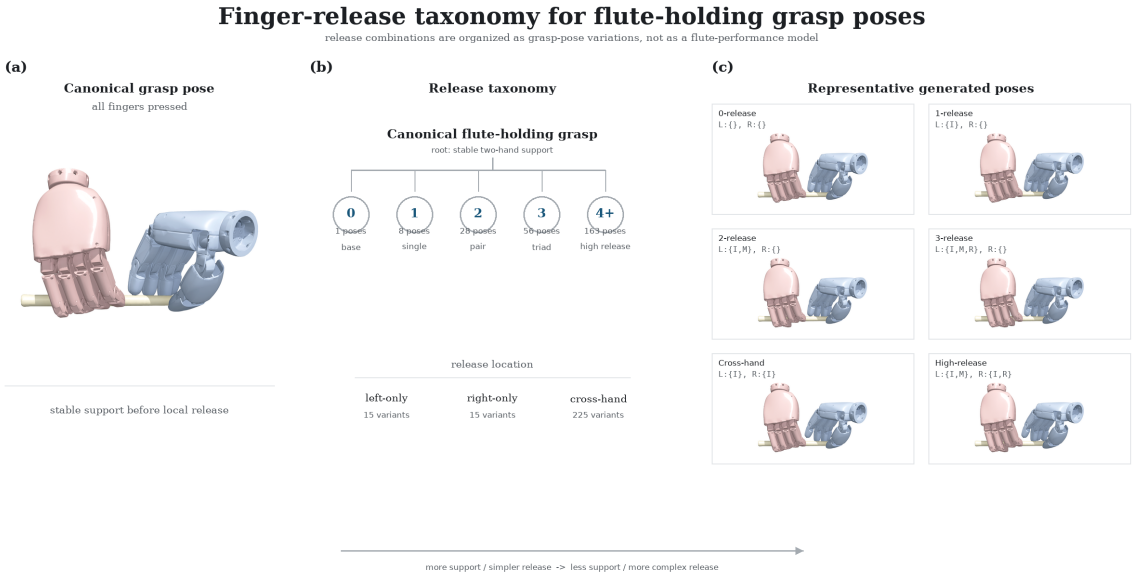

Figure 21. Finger-release taxonomy for flute-holding grasp poses. Starting from a canonical two-hand flute grasp, we organize generated poses by the number and location of released fingers. Compact labels such as $L{:}\{I, M, R\}, R{:}\{\}$ denote which fingers are released, where I, M, R, and P indicate index, middle, ring, and pinky fingers. The taxonomy provides an interpretable summary of the generated pose set and shows structured finger-level variation around a stable hand-object interaction; the flute context is used only as a task-motivated example of fine-grained dexterous grasp variation.

F Cross-Embodiment Study with the BODex Shadow Hand

The main paper evaluates SynManDex on UR5e+XHand because that is the physical embodiment used for trajectory generation and real-system validation. However, a stronger mechanistic test asks whether the human-prior proposal is tied to XHand, or whether it transfers to a different dexterous morphology. We therefore add a matched Shadow-hand study using the native BODex Shadow configuration [5]. This fixed-manifest study isolates one variable: *seed distribution*.

Hypothesis. If SynManDex uses MANO priors as geometry-aware pre-contact proposals, then a MANO pregrasp retargeted to the Shadow hand should improve the same BODex Shadow optimizer relative to its native random/geometric initialization, after controlling object mesh, object pose, scale, candidate budget, solver configuration, and evaluation metrics. The falsifier is equally clear: if the seeded variant does not improve force closure, penetration, contact consistency, or executable yield under this matched protocol, then the current human-prior mechanism is XHand-specific or loses useful contact geometry under the MANO→Shadow morphology gap.

Matched protocols. We compare two variants on the same GRAB object manifest:

- **BODex-Shadow random**: BODex's native right Shadow-hand grasp synthesis using `sim_shadow/fc.yml`, with its default seed generator and a fixed restart budget B.
- **SynManDex-Shadow seeded**: the same BODex Shadow solver and the same total restart budget B; the first restart is initialized by a MANO prior retargeted to the Shadow hand, and the remaining $B-1$ restarts use the same BODex seed generator as the baseline.

Both rows use the same object mesh $\mathcal{M}$, object pose $\mathbf{x}_o$, object scale, table configuration, collision geometry, optimizer iteration count, friction coefficient, and force-closure evaluator. Thus the only controlled difference is whether the candidate set contains a human-prior seed.

MANO-to-Shadow seed construction. Let $\mathbf{P}^H = \{\mathbf{p}_k^H\}_{k=1}^{11}$ be the MANO landmarks used for retargeting: wrist/palm, five distal finger joints, and five intermediate finger joints. Let ℓ_k be the corresponding Shadow links: `rh_palm`, distal links `rh_thdistal`, `rh_ffdistal`, `rh_mfdistal`, `rh_rfdistal`, `rh_lfdistal`, and middle links `rh_thmiddle`, `rh_ffmiddle`, `rh_mfmiddle`, `rh_rfmiddle`, `rh_lfmiddle`. We solve a position-retargeting problem over the floating Shadow root and the 22 actuated Shadow finger joints:

$$(\mathbf{T}_0^S, \boldsymbol{\theta}_0^S) = \arg\min_{\mathbf{T},\boldsymbol{\theta}} \sum_{k=1}^{11} \rho\Big(\Big\|\mathrm{FK}_{\ell_k}^S(\mathbf{T},\boldsymbol{\theta}) - \mathbf{p}_k^H\Big\|_2\Big) + \lambda_{\mathrm{rt}} \|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|_2^2, \tag{22}$$

where ρ is a Huber penalty and $\bar{\boldsymbol{\theta}}$ is the previous or neutral Shadow configuration. Before optimization, we align the MANO palm landmark frame to the Shadow zero-pose palm landmark frame; this avoids the common error of treating the local axes of `rh_palm` as if they were MANO axes. The resulting seed is

$$\mathbf{s}_{\mathrm{MANO}}^S = \left[\mathbf{t}_{\mathrm{palm}}^S, \mathbf{r}_{\mathrm{palm}}^S, \boldsymbol{\theta}_0^S\right] \in \mathbb{R}^{29}, \tag{23}$$

where $\mathbf{r}_{\mathrm{palm}}^S$ is represented as a unit quaternion in $wxyz$ order. BODex then runs its native Shadow-hand objective with either

$$\mathcal{S}_{\mathrm{rand}} = \{\mathbf{s}_1, \ldots, \mathbf{s}_B\} \quad \text{or} \quad \mathcal{S}_{\mathrm{human}} = \{\mathbf{s}_{\mathrm{MANO}}^S, \mathbf{s}_2, \ldots, \mathbf{s}_B\}, \tag{24}$$

holding B fixed.

Metrics. We report both retargeting diagnostics and final grasp quality. Retargeting diagnostics include mean and maximum landmark residual in Equation (22). Final grasp metrics mirror the main paper: number of valid grasps returned, force-closure rate, G1 stability, penetration depth, contact-region count, and, when the BODex UR10e+Shadow motion-generation path is enabled, arm-reachable trajectory success. A method-blinded VLM-H score can be computed from the same renderer used in Appendix H.2. The primary claim is physical and cross-embodiment transfer, with visual preference treated as a secondary diagnostic.

Visual diagnostics. Figure 17 gives the human-seeded Shadow-hand visual results under the matched protocol above. Table 9 reports the corresponding quantitative diagnostic. The remaining diagnostic in Figure 22 shows why the seed distribution matters for BODex-Shadow: random/geometric initialization can satisfy local hand-object contact while placing the wrist in configurations that make execution unusable.

Figure 22. BODex-Shadow failure cases. Panels 2 and 3 illustrate a common failure mode of random/geometric initialization: without a human prior, the optimizer can place the Shadow wrist into the table while pursuing local hand-object contact, producing wrist-table collision before a usable grasp configuration forms.

G Full Related Work

The condensed related work in Section 2 is expanded here. The first two subsections cover human priors, task semantics, and grasp taxonomies; the last two cover dexterous manipulation, grasp synthesis, and demonstration data.

G.1 Human Priors and Hand-Object Interaction

The parametric MANO model [16] and datasets, including GRAB [17], ContactPose [24], and DexYCB [25], encode rich distributions of natural human grasps. Generative models trained on these resources synthesize diverse hand-object interactions via diffusion [18, 27] and whole-body motion [19]. Broader motion diffusion models [42, 43] inspire the denoising architecture while operating in body-motion space instead of hand-object space. However, transferring these priors to robots faces two obstacles: the *morphological gap*, where MANO's DoFs versus actuated robot joints make direct retargeting infeasible [11] and existing bridging methods [12–14] do not enforce force closure. GeoRT [29] provides a cleaner retargeting principle by specifying geometric criteria such as local motion preservation, robot keypoint-space coverage, flat response, pinch correspondence, and collision avoidance. DexHiL [44] further highlights that dexterous systems should treat arm and hand mappings as coupled but distinct channels, since high-DoF hand motion and arm end-effector motion fail in different ways during data collection and correction. Joint motion manifold learning [45] offers an alternative by mapping human hand trajectories to robot actions; it still requires task-specific paired data and does not guarantee physical validity. A second obstacle is the *representation gap*: human HOI models output MANO-space poses, while robot policies need arm-constrained trajectories. SynManDex addresses both by treating human priors as *initialization* for physics-based refinement and by extending GeoRT-style geometry to the final arm-hand IK stage (Figure 2).

G.2 Task-Conditioned Grasp Taxonomies and Executable Evaluation

Recent task-template systems such as Dexonomy study a complementary axis: selecting object- and task-conditioned grasp primitives before low-level matching. This line is useful as a taxonomy and baseline source. Its static wrench-boundary scores measure a different property from full arm-hand execution, so we separate task-template evidence from IK-resolved lift success, source provenance, and failure modes. A grasp can score well under a contact-template or wrench metric and still fail when the UR5e+XHand embodiment cannot reach the wrist pose or execute the lift trajectory.

G.3 Dexterous Manipulation with Robot Hands

Dexterous manipulation with multi-fingered hands remains difficult because control must handle high-dimensional configuration spaces, contact-rich dynamics, and underactuation [1]. Reinforcement learning has shown impressive per-task results [2, 28, 46–49], with billions of interactions and task-specific reward engineering; bridging simulation to reality further demands careful demonstration augmentation [50]. Augmenting policy learning with demonstrations improves efficiency [51], yet demonstration quality is the bottleneck [52]: real multi-fingered demonstrations are prohibitively expensive, and existing synthetic generators produce only static floating-hand poses. Expressive trajectory architectures [53–55] are increasingly capable; data quality is now the limiting factor more often than model capacity. Neural tracking controllers trained on human references [56] further show that demonstration

quality is a bottleneck for generalizable dexterous control. Recent bimanual systems sharpen this point from complementary directions: UltraDexGrasp scales synthetic grasp-and-trajectory data and trains a point-cloud policy for universal grasping [22], while DexImit converts monocular human videos into physically plausible robot data through reconstruction, scheduling, action synthesis, and augmentation [15]. SynManDex targets this gap with a pipeline that converts human grasp priors into physically valid, arm-constrained bimanual trajectories (Section 3.2).

G.4 Dexterous Grasp Data Generation

Force-closure optimization remains central to dexterous grasp synthesis, with representative methods including DexGraspNet [3], BODex [5], and bimanual extensions [20, 34, 57]. Learning-based alternatives improve generalization via CVAEs [30], diffusion models [31], transformers [32], and geometry-aware curricula [58], while earlier generative work explored VAEs [59], GANs [60], implicit representations [61], and contact reasoning [33, 62]. Recent work extends grasp synthesis to sequential multi-object settings via diffusion [63]; the three shortcomings below remain open. DexGraspNet further evaluates grasp diversity using mean joint-angle entropy [3]; this supports comparability yet remains insufficient as a standalone metric, since high entropy can be produced by joint-limit, non-contact, or visually unnatural poses. These methods share three limitations: **(i)** random initialization converges to unnatural local minima, producing finger configurations no human would choose; **(ii)** all output static floating-hand poses, leaving the mismatch to arm-constrained execution unresolved despite being identified [5]; **(iii)** static poses are insufficient as imitation-learning demonstrations, which require multi-phase approach-grasp-lift trajectories. SynManDex addresses all three via human-prior initialization (**i**; Section 3.2), GeoRT-guided IK (**ii**; Section 3.2), and phase-structured trajectory generation (**iii**; Section 3.2).

H Full Experiment Protocols

Section 4 is written as a staged argument from mechanism-level validation to executable simulation and hardware verification. This appendix gives the corresponding evidence controls: fixed manifests, thresholds, baseline wrappers, trial definitions, held-out splits, and evaluator formulas. The order mirrors the main text: static grounding and human-likeness diagnostics, taxonomy and benchmark comparisons, policy ablations, manipulation trials, hardware validation, and the cross-embodiment Shadow-hand diagnostic.

H.1 Main-Paper Experiment Protocols

All comparisons in Section 4 use fixed object manifests and fixed evaluator thresholds for every method in the same table. Unless a row is explicitly marked as pose-only, a sample is counted as successful only after passing the full typed chain: contact feasibility, collision filtering, force-closure score, arm-hand IK, and a 10 cm lift or task-specific terminal-possession check. This convention is the link between the main figures and the quantitative tables: qualitative images show representative accepted or failed cases, while the reported numbers require the full execution gate.

Static grasp refinement. Table 2 is computed on the 312-object, 25-class manifest used for the main human-likeness and lift-validity claim in Section 4.2. Each method receives the same 240 candidate budget per object and the same object poses. Retarget-only uses the GeoRT-calibrated MANO-to-XHand output with no physical refinement. Optimization-only starts from random XHand wrist and joint seeds, then runs the same force-closure objective and collision terms as SynManDex. SynManDex starts from the retargeted human-prior seed and optimizes with the same budget and termination criteria. G1 is the Ferrari–Canny wrench-space margin under the contact model in Equation (5); penetration is the maximum hand-object interpenetration depth; Contact requires at least three active contact regions; FC requires $Q_{\mathrm{FC}} > 0$.

Human-likeness and diversity. Tables 2 and 10 share the same rendered grasp set. The VLM evaluator scores every retained method output under method-blinded views and a fixed JSON rubric; the human audit scores a stratified method-blinded subset with the same factors. The combined score is $0.6H^{\mathrm{VLM}} + 0.4H^{\mathrm{human}}$. PCD uses the human-manifold coverage and weighted-DPP protocol in Appendix H.2, so random high-entropy postures do not increase diversity unless they remain physically valid and human-like.

Taxonomy and benchmark comparisons. We distinguish matched comparisons from contextual comparisons. Matched comparisons use the same object-task cells, XHand collision model, candidate bud-

get, and execution gates. Contextual comparisons report the closest available protocol when a method is tied to its own task taxonomy, embodiment, or validation pipeline. For Table 3, Dexonomy-XHand maps selected taxonomy templates to XHand joint targets before applying the same contact, force-closure, IK, and lift filters as SynManDex. BODex, dex-retargeting, and GeoRT rows are run with matched object-task cells and the same XHand collision model. Unimanual rows evaluate one XHand; bimanual rows evaluate paired XHands plus UR5e reachability and lift. For Table 4, all baselines are evaluated with the same standardized static grasp checks: penetration, force closure, and benchmark success under the matched contact model.

Trajectory-policy ablations. Tables 5, 11 and 12 use the same held-out object split, policy backbone, action horizon, and 10 cm lift threshold. Data-source ablations change only the demonstration source used to train the policy. Policy-interface ablations keep the demonstration set fixed and remove either robot points or action-query readout. The open-loop row in Table 11 is a diagnostic replay upper bound; all policy rows are closed loop and re-observe point clouds after each action chunk.

Prehensile manipulation applications. Tables 6 and 7 evaluate whether the optimized keyframes remain useful after the initial lift stage in Section 4.5. In-grasp trials are counted successful when the object stays in contact through the commanded transition, slips by less than the reported threshold, and passes the final force-closure/lift check. Self-handover trials start from a two-hand passing keyframe and count success only if the releasing hand opens without losing the object. Pick-and-place trials add a target-zone constraint after lift; failures are categorized by the first violated condition.

Real-system protocol. The physical platform is a bimanual UR5e+XHand system with two calibrated Azure Kinect DK cameras observing a tabletop workspace (Figure 14). Table 8 uses three everyday objects with 10 physical trials each: vase, apple, and spray bottle. Before each trial, the object is reset to the same tabletop region, the fused point cloud is re-acquired, and the policy replans in the same 36-DoF arm-hand command space as simulation. A trial is successful when the robot establishes stable contact, lifts or transports the object to the task endpoint, and maintains possession at the terminal check. Lift valid is recorded separately because a rollout can momentarily lift the object yet fail final task completion due to slip, rolling, or placement error. The comparison rows in Table 8 use the same 30 physical trials and reset protocol as the main result. These trials validate executable transfer on the physical bimanual platform under a controlled tabletop protocol; they are not intended as a full real-world benchmark over object categories.

Extended real-world functional trials. Figure 16 is a qualitative hardware stress test rather than an additional row in Table 8. Its role is to connect the manipulation evidence in Section 4.5 with hardware execution under object and task variation not covered by the three-object success table. Camera lifting uses an additional toy camera outside the three-object hardware table to illustrate whether the action-chunk policy and lift gate can be applied to object instances beyond the quantitative hardware benchmark. Pick-handover-put follows the same staged structure as the simulated handover and pick-and-place protocols in Figures 12 and 13: pickup, transfer, release, and placement. Pouring uses a tilted terminal state to illustrate whether the grasp remains stable while expressing an object function rather than merely maintaining a vertical lift. The frames in Figure 16 are selected from the recorded videos to show the relevant state transitions; the project page provides the corresponding video clips.

H.2 Diversity and Human-Likeness Protocol

Entropy alone is not the right diversity objective for human-like dexterous grasping. Following DexGraspNet [3], we retain mean joint-angle entropy for comparability and use *Plausibility-Constrained Diversity* (PCD) as the main diversity metric. Let $\boldsymbol{\phi}_i$ be an object-normalized grasp descriptor containing wrist position/orientation, approach direction, normalized joint angles, fingertip distances, and contact-pattern features. Let $\mathcal{R} = \{\mathbf{r}_m\}_{m=1}^{M}$ be reference anchors from held-out GRAB/MANO grasps and retargeted human-prior seeds that are not used as generated outputs in the evaluated set. The reference set is fixed before evaluating generated grasps and is not updated using accepted samples from any compared method. Each generated grasp receives a soft plausibility weight

$$w_i = \sigma\left(\frac{Q_i - \tau_Q}{T}\right)\sigma\left(\frac{H_i - \tau_H}{T}\right), \tag{25}$$

Table 10. Diversity and human-likeness diagnostic. HMC and PCD are normalized to $[0, 1]$; VLM-H, Human-H, and Comb.-H are on a 1–5 scale; lift validity uses the main-text 10 cm threshold.

	Diversity / plausibility			Human-likeness			Execution
Method	Entropy ↑	HMC ↑	PCD ↑	VLM-H ↑	Human-H ↑	Comb.-H ↑	Lift valid (%) ↑
Retarget-only	2.18	0.44	0.09	4.28	4.03	4.18	8.1
Optimization-only	**2.84**	0.31	0.11	2.86	2.74	2.81	35.4
SynManDex full	2.61	**0.72**	**0.41**	**4.72**	**4.59**	**4.67**	**65.8**

where Q_i is physical quality, H_i is human-likeness, and T is a temperature. We compute human-manifold coverage

$$\mathrm{HMC} = \frac{1}{M}\sum_{m=1}^{M}\left[1 - \prod_i \left(1 - w_i \exp\left(-\frac{\|\boldsymbol{\phi}_i - \mathbf{r}_m\|_2^2}{2\sigma_R^2}\right)\right)\right], \tag{26}$$

and weighted DPP distinctness

$$\mathrm{DPP} = \frac{1}{N}\log\det\left(\mathbf{I} + \mathbf{W}^{1/2}\mathbf{K}\mathbf{W}^{1/2}\right), \qquad K_{ij} = \exp\left(-\frac{\|\boldsymbol{\phi}_i - \boldsymbol{\phi}_j\|_2^2}{2\sigma^2}\right), \tag{27}$$

where $\mathbf{W} = \mathrm{diag}(w_i)$. The final score is

$$\mathrm{PCD} = \mathrm{HMC} \cdot \mathrm{DPP}. \tag{28}$$

Thus a random high-entropy posture contributes little if it is off the human manifold or fails physical checks, while multiple distinct, plausible grasp strategies increase the score.
For H_i, we use a method-blinded VLM-based perceptual diagnostic as an independent human-likeness measure, separate from the regularization term in Equation (4). This score is used only for human-likeness analysis; physical success is determined by the contact, collision, force-closure, IK, and lift or task-completion gates. Each grasp is rendered under a fixed grid of views; Qwen2.5-VL [64] is the default scorer and DeepSeek-VL2 [65] is used on a 10–20% audit subset. The prompt asks for JSON-only scores on functional contact, ergonomic approach, finger coordination, object appropriateness, artifact penalty, and confidence; the normalized score is

$$H_i^{\mathrm{VLM}} = \frac{.25h_{\mathrm{contact}} + .20h_{\mathrm{approach}} + .25h_{\mathrm{coord}} + .20h_{\mathrm{object}} + .10(5 - h_{\mathrm{artifact}})}{5} \cdot \left(.5 + .5\frac{h_{\mathrm{conf}}}{5}\right). \tag{29}$$

We calibrate this score with positive human-prior anchors and negative corruptions, then bootstrap confidence intervals across objects. For the human audit, annotators score the same rendered grids using the identical five-factor rubric without method labels. The headline human-likeness score combines the scalable VLM pass with the human audit:

$$H_i^{\mathrm{comb}} = 0.6\, H_i^{\mathrm{VLM}} + 0.4\, H_i^{\mathrm{human}}. \tag{30}$$

Qualitative diversity analysis. Figure 9 visualizes the qualitative side of PCD. The accepted samples are diverse along object-relative contact, hand-role assignment, and approach direction rather than only along joint-angle spread. Across the six examples, one hand may stabilize a broad surface while the other pinches a rim, two hands may cage a thin object from opposite sides, or both hands may share support on a tall object. This is the intended behavior of plausibility-constrained diversity: the dataset should cover multiple functional contact modes while staying inside the physically valid, human-like region counted by Table 10.

H.3 Policy and Executable-Mapping Protocols

Table 11 specifies controlled representation ablations for the point-cloud policy. The main comparison is closed-loop execution under fresh observations and held-out objects, not offline reconstruction.
Because GeoRT-IK replaces the earlier wrist-only IK layer, we include a matched executable-mapping ablation that holds the optimized grasp pool, object poses, candidate budget, and lift threshold fixed.

Table 11. Policy protocol. Seen and heldout report rollout success (%); drift is terminal joint-space error during executable rollout. The open-loop replay row is a diagnostic reference and is not included in the closed-loop policy ranking.

Protocol	Observation	Inference	Seen (%)	Heldout (%)	Drift
Open-loop diagnostic	Scene+robot point cloud	Full replay	96.4	93.2	0.118
Scene-only policy	Scene point cloud	Closed-loop	54.3	45.7	0.539
Pooled-feature policy	Scene+robot point cloud	Closed-loop	51.4	40.0	0.601
Full point-cloud policy	Scene+robot point cloud	Closed-loop	**88.9**	**80.7**	0.474
Image-token baseline	Fixed dual RGB-D	Closed-loop	48.6	36.4	0.647

Table 12. Matched GeoRT-IK ablation. All rows use the same optimized grasp candidates and SAPIEN lift evaluator. Wrist/Kpt errors are terminal residuals; IK valid and Lift are pass rates.

	Residual		Validity		Runtime
Executable mapping	Wrist err. (cm/deg) ↓	Kpt err. (cm) ↓	IK valid (%) ↑	Lift (%) ↑	Time (s) ↓
Wrist-only executable mapping	**0.7/2.8**	5.9	42.1	31.4	**0.42**
LM hand seed + wrist IK	0.8/3.1	4.8	48.6	36.2	0.51
GeoRT hand seed + wrist IK	**0.7/2.9**	3.1	56.7	45.8	0.54
Full GeoRT-IK (ours)	0.9/3.4	**1.2**	**82.3**	**65.8**	0.76

I Domain Randomization

For the point-cloud policy, visual and material domain randomization improves perception under rendering variation. We randomize the following parameters independently per episode:

Table 13. Domain randomization parameters applied during trajectory collection.

Parameter	**Distribution**	**Range**
Object diffuse color	Uniform HSV	$H \in [0,1],\ S \in [0.3,1],\ V \in [0.3,1]$
Table surface color	Uniform RGB	$[0,1]^3$
Table material roughness	Uniform	$[0.3,\ 0.9]$

J Point-Cloud Policy Architecture Details

The policy follows a simple point-cloud control pattern: encode scene geometry, aggregate it with action query tokens, and predict a bounded action distribution. We use this architecture because the dataset is generated from 3D meshes and executed in a simulator with known robot state; point clouds preserve the object, hand, and table geometry in a common metric frame.

J.1 Point Encoder and Action Queries

The input point cloud is downsampled by farthest-point sampling to a fixed budget. A PointNet++-style encoder [40] extracts local geometric features with set abstraction layers. Learnable action query tokens attend unidirectionally to the point features and produce a chunk of action latents. The same readout is used across approach, contact, and lift phases; the current scene and robot geometry carry the phase information.

J.2 Training Details

Table 14. Point-cloud policy training configuration.

Hyperparameter	Value
Input points	2048
Point encoder	PointNet++ set abstraction
Attention readout	action queries with unidirectional attention
Action distribution	truncated normal
Chunk size H	16
Control dimension	36
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	1×10^{-4}
Gradient clipping	1.0
Epochs	50
Mixed precision (AMP)	enabled

The training loss is the negative log-likelihood in Equation (11). Because the action distribution is truncated to joint limits, the model cannot assign probability mass to commands outside the executable control range.

J.3 Inference

At test time, the policy predicts an action chunk and executes the first short horizon before observing a fresh point cloud and replanning. This receding-horizon strategy allows feedback correction while still training on temporally coherent demonstration windows.

K Limitations

Our evaluation focuses on controlled bimanual dexterous grasp synthesis, executable mapping, and hardware validation on a tabletop bimanual system. PCD and VLM-H are used as diagnostic measures of diversity and human-likeness rather than as standalone physical-success criteria; physical validity is still determined by the contact, collision, force-closure, IK, and lift or task-completion gates used throughout the evaluation. To reduce evaluator bias, human-likeness scores are computed under method-blinded views and calibrated with a human audit.

Cross-benchmark comparisons are interpreted under their matched evaluation scope. When object poses, candidate budgets, collision models, and execution gates can be aligned, we report direct matched comparisons; when external methods rely on different task templates or validation protocols, we use them as contextual references rather than strict one-to-one replacements. The Shadow-hand experiment is therefore presented as a fixed-manifest embodiment-transfer diagnostic, not as a complete Shadow-hand benchmark.

Remaining failures are concentrated around embodiment-specific reachability, palm-support loss during retargeting, and dynamic contact robustness during squeeze-and-lift execution. These observations motivate future proposal models that jointly account for object geometry, hand morphology, and execution-time contact stability.