

ObjRetarget: An Object-Aware Motion Retargeting Framework with Anthropomorphic Arm Constraints and Polyhedral Hand Modeling

Yuanchuan Lai¹, Qing Gao^{1,*}, Ziyang Liang¹, Junjie Hu², Zhaojie Ju³

Abstract—Learning robot dexterous manipulation from human manipulation videos requires reliably retargeting human intent to executable robot actions while maintaining stable hand–object contact, which remains a key challenge in embodied intelligence. Existing retargeting methods often ignore explicit contact modeling or rely on reinforcement learning, resulting in limited accuracy and generalization. To address this, we propose ObjRetarget, a human-to-robot motion retargeting framework for learning robot dexterous manipulation from human videos, which integrates anthropomorphic arm trajectory constraints with structured hand–object geometric modeling. For arm motion, reference trajectories extracted from human videos are used for initialization, followed by anthropomorphic constraints and redundancy-aware optimization to generate natural and accurate movements. For hand manipulation, ObjRetarget represents multi-finger contacts using polytope clusters and preserves contact structure through geometric invariants to improve stability. Experiments on real robots show that ObjRetarget improves manipulation success rates and contact stability across multiple dexterous tasks, and generalizes well to different demonstrations, object poses, and task settings. Project can be found at: <https://github.io/ObjRetarget>.

I. INTRODUCTION

Learning robot dexterous manipulation from human manipulation videos requires reliably retargeting human intent into executable robot actions while maintaining stable hand–object contact throughout the manipulation process. This remains a central challenge in embodied intelligence and dexterous manipulation research, as successful manipulation depends not only on motion imitation but also on preserving physically meaningful contact relationships [1]–[3]. Existing motion retargeting approaches mainly focus on pose or joint mapping in free space, where the robot is encouraged to reproduce demonstrated kinematics [4], [5], as illustrated in Fig. 1 (a). However, modeling object exchange, contact transitions, and sustained hand–object interaction is still limited, particularly in approaches that rely on direct hand mapping [6], [7] (Fig. 1 (b)). In real-world manipulation, robots must not only reproduce fine-grained finger

This work was supported in part by the the Shenzhen Science and Technology Program under Grant ZDCY20250901100201002, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2025A1515011954, 2023A1515110074.

¹Yuanchuan Lai, Qing Gao, Ziyang Liang are with the School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen 518107, China.(email: laiych25@mail2.sysu.edu.cn, gao-qing2@mail.sysu.edu.cn)

²Junjie Hu is with the School of Artificial Intelligence, The Chinese University of Hong Kong, Shenzhen, Shenzhen 518172, China.(email: hujunjie@cuhk.edu.cn)

³Zhaojie Ju is with the School of Computing, University of Portsmouth, Portsmouth PO1 3HE, UK.(email: Zhaojie.Ju@port.ac.uk)

*Corresponding Author: Qing Gao, gaoqing2@mail.sysu.edu.cn.

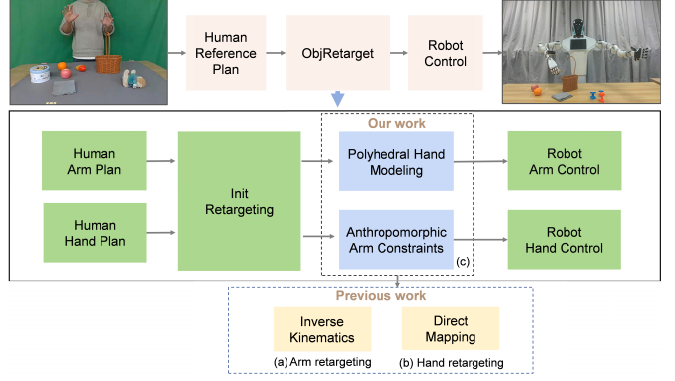


Fig. 1. Schematic of the motion retargeting framework. Part (a) shows arm retargeting via inverse kinematics, part (b) shows hand retargeting via direct mapping, and part (c) presents the proposed pipeline with polyhedral hand modeling and anthropomorphic arm constraints.

motions but also coordinate large-scale arm movements, ensuring that grasping, transport, and object operation remain both physically stable and behaviorally natural. This requires reasoning about contact geometry, motion continuity, and task-dependent constraints simultaneously.

However, most existing methods treat the arm and hand as a unified system, overlooking their fundamental differences in motion scale, functional roles, and precision requirements [8]. The arm primarily produces large, global movements that determine object approach, spatial positioning, and overall motion trajectory, whereas the hand performs fine, contact-sensitive actions that directly govern grasp stability and manipulation success. Ignoring this distinction often leads to conflicts between global motion smoothness and local contact stability, making it difficult to simultaneously achieve robustness and naturalness in tasks such as object transfer, contact establishment, and contact maintenance. Meanwhile, many interaction-aware approaches rely on reinforcement learning for end-to-end optimization [9], [10]. Although such methods can implicitly learn contact behaviors, their generalization ability is often limited, and performance can degrade when object geometry, initial poses, or task conditions vary. These limitations suggest that stable and generalizable retargeting requires not only reproducing demonstration motions, but also explicitly modeling hand–arm motion differences and incorporating object interaction constraints into the retargeting process.

To address these challenges, we propose ObjRetarget, an object-interaction-oriented human-to-robot retargeting framework for dexterous manipulation, as illustrated in

Fig. 1. The pipeline first builds a human reference plan by extracting articulated joint trajectories and object point-cloud information from demonstration videos, followed by an initial retargeting stage that retargets human motion into a coarse robot trajectory for subsequent optimization. Building upon this initialization, ObjRetarget explicitly models the distinct motion characteristics of the arm and hand. For arm motion, we propose to refine the retargeted reference trajectories through anthropomorphic motion constraints and task-adaptive optimization, enabling the global motion to follow human movement patterns while adapting to variations in object initial poses, thereby improving motion naturalness and accuracy. For hand manipulation, we propose to construct polytope clusters in which each finger-object contact forms a local geometric unit, and geometric invariant constraints are applied to preserve contact structure and manipulation semantics, enhancing grasp stability and interaction reliability. Extensive real-robot experiments show that ObjRetarget significantly improves manipulation success rates and contact stability across multiple dexterous manipulation tasks, while maintaining natural and interpretable robot motions.

Our contributions are summarized as follows:

- We propose ObjRetarget, a hand-arm decoupled human-to-robot retargeting framework that explicitly separates global arm motion from contact-sensitive hand manipulation, enabling object-aware motion generation within a unified pipeline.
- We develop a reference-trajectory-guided arm optimization strategy with anthropomorphic motion constraints, producing stable and human-consistent arm trajectories under varying object poses and task conditions.
- We introduce a polytope-cluster hand modeling method that represents multi-finger contacts as local geometric units and preserves hand-object contact structure through geometric constraints.

II. RELATED WORK

A. Object-Aware Dexterous Manipulation

Object-aware dexterous manipulation is challenging due to multi-contact constraints and complex hand-object geometry. Reinforcement learning and model predictive control are widely used to learn interaction policies [11], [12], but reinforcement learning typically requires extensive training processes with large-scale interaction rollouts and careful reward design, and its performance can be sensitive to variations in object geometry and initial conditions [13], [14]. Other approaches incorporate explicit contact models or tactile sensing to improve stability [15], [16], yet they often depend on accurate contact modeling or specialized tactile hardware, limiting scalability and cross-object generalization. As a result, achieving stable and generalizable dexterous manipulation without heavy training procedures or precise object modeling remains difficult.

To overcome these limitations, ObjRetarget leverages human manipulation priors from videos rather than relying on autonomous policy learning. It models hand-object interaction using polytope clusters with geometric invariants

and guides arm motion with reference trajectories from demonstrations. By avoiding large-scale training procedures and explicit object models, this structured, human-informed formulation enables stable interaction and improved generalization across diverse objects and task scenarios.

B. Human-to-Robot Motion Retargeting

Human-to-robot motion retargeting transfers human demonstrations to robot platforms with different kinematics and degrees of freedom for high-fidelity motion reproduction. Traditional methods rely on analytical kinematics, optimization, or constraint-based planning, typically using joint-space mapping to adapt arm and hand poses [17]–[19], while more recent works employ deep learning or latent representations to reduce kinematic discrepancies and improve retargeting accuracy [20]–[22]. These approaches perform well in free-space imitation and trajectory tracking.

However, most methods assume object-free or weak-contact settings and focus on geometric consistency or joint error minimization, without explicitly modeling hand-object interactions. In addition, many approaches treat the arm and hand as a unified system [23], [24], overlooking their distinct roles: the arm controls large-scale motion, while the hand requires precise, contact-sensitive manipulation. This simplification can reduce coordination between global motion and local interaction, degrading performance in complex tasks.

ObjRetarget addresses these limitations by separating arm-level global motion from hand-level fine manipulation. Arm motion is guided by reference trajectories and refined with task-related constraints, producing physically executable motions that preserve human characteristics and support stable object interaction. At the same time, the retargeting process explicitly accounts for hand-object contact, ensuring stable and precise multi-finger manipulation during interaction.

C. Learning from Human Demonstrations

Human demonstration videos provide rich priors for robot dexterous manipulation, enabling object-aware motion retargeting from humans to robots. Traditional imitation-learning methods map human videos or motion capture data to robot control policies through trajectory replay, behavior cloning, or end-to-end visuomotor learning [25]–[28]. While effective for skill acquisition, these approaches typically focus on autonomous policy learning and often ignore fine-grained hand-object interactions, limiting precise motion mapping and stable grasp execution.

Single-video retargeting methods, such as OKAMI [31] and ORION [32], leverage human demonstrations to generate executable robot motions. OKAMI imitates humanoid motion for skill learning but struggles with multi-stage operations and continuous contact control due to the lack of explicit hand-object modeling. ORION uses open-world object graphs to infer object-centric actions, improving generalization across objects and layouts, but does not model fine-grained hand-object contacts. These examples highlight the potential of video-driven manipulation while showing the

need for structured, object-aware retargeting. ObjRetarget addresses this by modeling hand–object contacts with polytope clusters and guiding arm motion via human reference trajectories, enabling contact-consistent, anthropomorphic, and generalizable motion suitable for downstream skill learning.

III. OBJRETARGET

A. Object-Aware Motion Retargeting Framework

We propose ObjRetarget, an object-interaction-oriented motion retargeting framework that learns robot manipulation from human operation videos. As illustrated in Fig. 2 (a), given an RGB-D manipulation video, human body motion is first reconstructed using the SLAHMR [29] pose estimator, an iterative optimization-based algorithm for recovering human motion sequences. The manipulated objects are then identified using a vision–language model, followed by object pose tracking to estimate their spatial trajectories over time. These components are integrated to obtain synchronized sequences of human joint poses, object point-cloud poses, and human–object interaction distances. Contact events are detected by thresholding the interaction distance, enabling the system to distinguish between free motion and physical interaction phases.

At the arm level (Fig. 2 (b)), ObjRetarget generates global motion through a two-stage strategy. Human joint pose sequences are first used to produce an initialization retargeting trajectory, providing a feasible and structurally consistent motion prior. This trajectory is then refined by incorporating object pose information together with anthropomorphic motion constraints, yielding an executable arm trajectory that remains consistent with the demonstration while adapting to specific task configurations. At the hand level (Fig. 2 (c)), ObjRetarget adopts a contact-aware staged retargeting mechanism. During the non-contact phase, the system follows the initialized hand motion obtained from retargeting. Once contact is detected, polytope-cluster-based geometric consistency optimization is activated to preserve local contact structures and manipulation semantics.

Finally, Fig. 2 (d) shows the unified temporal scheduler responsible for coordinating arm and hand trajectories. By synchronizing global arm motion with fine-grained hand manipulation, ObjRetarget ensures coherent dual-arm execution. Leveraging this hand–arm decoupled formulation, the framework achieves natural, stable, and generalizable dexterous manipulation in complex object-interaction tasks.

B. Anthropomorphic Arm Constraints

To obtain stable, natural, and generalizable robot execution trajectories in complex object-interaction tasks, ObjRetarget adopts a two-stage arm motion generation strategy. First, a reference trajectory is generated based on our previous retargeting work [30], providing a feasible and structured initial motion. Then, the trajectory is refined using an optimization framework guided by anthropomorphic motion constraints and task objectives, producing motions that satisfy task requirements while maintaining anthropomorphic motion patterns.

Motion Initialization via Motion Retargeting: Although the initialization trajectory may not be the final execution result, it provides a structurally valid and near-optimal starting point, which greatly improves convergence stability in subsequent optimization.

Let the human motion sequence, consisting of T frames, be represented as a set of skeleton graphs $D = \{G_k\}_{k=1}^T$, where each graph $G_k = (V_k, E_k, W_k)$ consists of joint nodes V_k , skeletal connectivity E_k , and dynamic attention weights W_k at frame k . Each skeleton graph is embedded into a latent space via an encoder $f_\psi(\cdot)$, producing a compact representation $z_k = f_\psi(G_k)$, which is then decoded into the robot joint configuration $\theta_k = f_\phi(z_k) \in R^n$, where n is the number of robot joints and θ_k respects the joint limits.

The corresponding robot end-effector pose S_k is obtained via forward kinematics, $S_k = FK(\theta_k)$, where S_k encodes both position and orientation of the end-effector. The retargeting process is optimized by minimizing a reference loss:

$$L_{ref}(S_k, S_k^h) = L_{pos} + \lambda_{ori} L_{ori} + \lambda_{tip} L_{tip}, \quad (1)$$

which accounts for end-effector position error L_{pos} , orientation error L_{ori} , and fingertip consistency L_{tip} relative to the human demonstration S_k^h , with λ_{ori} and λ_{tip} weighting the contributions. The optimization over encoder and decoder parameters (ψ, ϕ) is formulated as:

$$\min_{\psi, \phi} \sum_{k=1}^T L_{ref}(S_k, S_k^h), \quad (2)$$

$$s.t. \quad \theta_{\min} \leq \theta_k \leq \theta_{\max}. \quad (3)$$

The resulting robot trajectory $S = \{S_k\}_{k=1}^T$ serves as the macro-level arm reference, providing an effective initialization for subsequent geometric consistency optimization.

Task-Adaptive Arm-Plane Regularization: In human motion, the plane formed by the shoulder, elbow, and wrist evolves smoothly over time, and its normal direction is closely related to motion direction and task semantics. In contrast, robots with redundant degrees of freedom often produce non-anthropomorphic elbow configurations when only end-effector constraints are enforced, including elbow flipping, jitter, or abnormal lifting.

Existing approaches typically promote anthropomorphic motion by directly matching the arm-plane normal of the human demonstration. However, this strategy effectively treats the arm plane as a rigid pose target, which restricts motion flexibility in constrained manipulation scenarios. To maintain anthropomorphic structure while allowing adaptive motion, the arm plane is instead formulated as a soft geometric constraint.

As shown in Fig. 3, let the robot shoulder, elbow, and wrist positions at time t be $p_s(t)$, $p_e(t)$, $p_w(t)$. The instantaneous arm-plane normal is defined as:

$$n_{arm}^{robot}(t) = \frac{(p_e(t) - p_s(t)) \times (p_w(t) - p_e(t))}{\|(p_e(t) - p_s(t)) \times (p_w(t) - p_e(t))\|}. \quad (4)$$

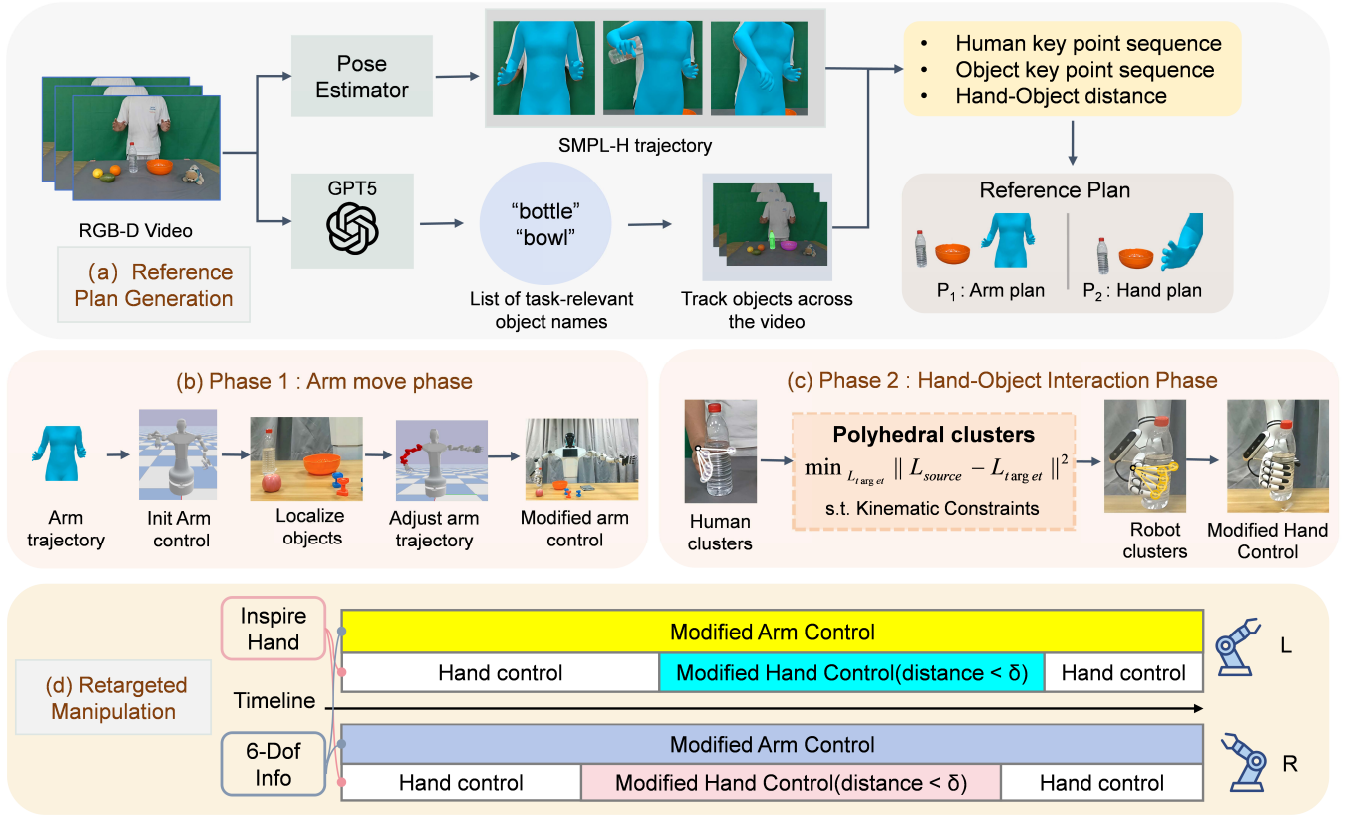


Fig. 2. Overview of the object-aware motion retargeting framework. Part (a) illustrates the extraction of human 3D body and hand poses from an RGB-D video, along with the detection and tracking of task-relevant objects to generate a reference plan. Part (b) presents arm trajectories refined based on object poses and anthropomorphic constraints. Part (c) shows hand motions optimized at contact points using polyhedral clusters. Part (d) depicts a unified temporal scheduler that synchronizes arm and hand motions for coordinated dual-arm execution.

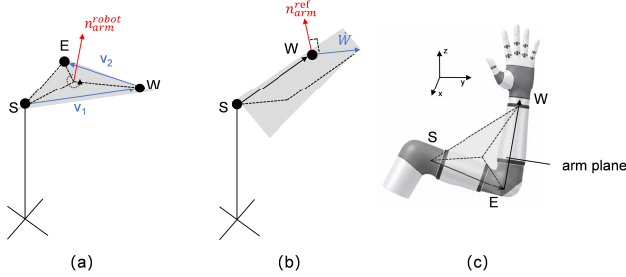


Fig. 3. Schematic of the task-adaptive arm-plane regularization. (a) The normalized normal vector $\mathbf{n}_{arm}^{robot}$ defined by the shoulder-elbow-wrist triangle. (b) The reference normal $\mathbf{n}_{arm}^{ref}$ constructed from the wrist motion direction. (c) The arm plane and its normal vector in the robot configuration.

Instead of using the human demonstration directly, the reference bending direction is constructed adaptively from the current motion trend:

$$\mathbf{n}_{arm}^{ref}(t) = \frac{(p_w(t) - p_s(t)) \times \dot{p}_w(t)}{\|(p_w(t) - p_s(t)) \times \dot{p}_w(t)\|}, \quad (5)$$

where $\dot{p}_w(t)$ denotes the wrist velocity direction. The resulting task-adaptive arm-plane loss is defined as:

$$L_{plane}(t) = w(t) (1 - \mathbf{n}_{arm}^{robot}(t) \cdot \mathbf{n}_{arm}^{ref}(t)), \quad (6)$$

This adaptive weighting $w(t)$ emphasizes the prior during

large, fast arm motions while automatically reducing it during fine contact manipulation, preventing interference with task accuracy. The formulation penalizes only large deviations from human-consistent bending trends while preserving redundancy flexibility.

Task-Oriented End-Effector Tracking Loss: To ensure task accuracy during trajectory refinement, ObjRetarget includes a standard end-effector pose tracking objective. Let the wrist pose at time t be $x_w(t) = [p_w(t), R_w(t)]$, where $p_w(t) \in \mathbb{R}^3$ is the wrist position and $R_w(t) \in SO(3)$ is the wrist orientation represented as a rotation matrix. Similarly, the reference wrist pose is $x_w^{ref}(t) = [p_w^{ref}(t), R_w^{ref}(t)]$, where $p_w^{ref}(t)$ and $R_w^{ref}(t)$ denote the target position and orientation at frame t . The task-oriented tracking loss is defined as:

$$L_{task}(t) = \|p_w(t) - p_w^{ref}(t)\|^2 + \lambda_R \|\phi(R_w^{ref}(t)^\top R_w(t))\|^2, \quad (7)$$

where $\phi(\cdot)$ maps the rotation error from $SO(3)$ to its Lie algebra representation, and λ_R weights the orientation term relative to the position term. This loss focuses purely on task accuracy and does not impose any additional priors, ensuring generality across different scenarios.

Overall Optimization Objective: Combining task constraints and anthropomorphic regularization, the arm trajec-

tory optimization is formulated as:

$$\min_{q(t)} \sum_t \left[L_{\text{task}}(t) + \lambda_p L_{\text{plane}}(t) + \lambda_s \|q(t) - q(t-1)\|^2 \right], \quad (8)$$

where $q(t)$ denotes the robot joint configuration at time t , and λ_p and λ_s control the anthropomorphic regularization and temporal smoothness terms, respectively.

C. Polyhedral Hand Modeling

Hand-object contact representation: To accurately capture fine-grained geometric relationships between the human hand and objects during multi-finger manipulation, ObjRetarget introduces an object-aware hand-object contact representation. Given an RGB-D input video, the system first extracts hand keypoints and the object surface point cloud. The hand keypoints include the five fingertips and the palm center (six points in total) to represent the overall hand configuration, while the object surface is reconstructed from the depth point cloud.

Based on the spatial adjacency between hand keypoints and object points, the system performs finger-wise contact detection to determine which fingers are in effective contact with the object. For a contacting finger f , its local contact region C_f on the object surface is extracted, and the centroid of this region is used as the representative contact point:

$$c_f = \frac{1}{|C_f|} \sum_{o \in C_f} o, \quad (9)$$

where $|C_f|$ denotes the number of points in the contact region, and o represents points on the object surface. This design enables robust extraction of multi-finger contact information under complex interaction scenarios, providing reliable input for structured geometric modeling.

Polytope cluster construction: After obtaining finger-wise contacts, ObjRetarget constructs polytope clusters to model local geometric relationships during multi-finger collaboration. For a contacting finger f , a local tetrahedral unit T_f is defined by four vertices: the palm center, the fingertip keypoint of f , the fingertip of an adjacent finger, and the corresponding object contact point c_f . When multiple fingers contact simultaneously, the resulting tetrahedra form a polytope cluster representing the local hand-object geometry under coordinated manipulation. This formulation transforms complex multi-finger contact relationships into a set of stable, compact geometric units, providing interpretable and optimizable high-dimensional contact constraints.

The core idea is that human manipulation does not require exact reproduction of absolute poses but rather the preservation of stable local geometry and contact semantics. Consequently, polytope clusters serve as soft geometric constraints, providing a structured basis for subsequent geometric consistency optimization.

Geometric-invariant optimization: Based on polytope-cluster modeling, ObjRetarget formulates hand motion retargeting as a nonlinear optimization problem under geometric

invariants. The overall geometric consistency loss is defined as:

$$L = \sum_f w_f \cdot L(T_f), \quad (10)$$

where f indexes the fingers, w_f is a weight coefficient, and $L(T_f)$ is the geometric consistency loss for the f -th tetrahedron, composed of edge-length and relative-pose invariants:

$$L(T_f) = L_{\text{edge}}(T_f) + \lambda L_{\text{pose}}(T_f), \quad (11)$$

The edge-length loss constrains the local tetrahedral geometry:

$$L_{\text{edge}}(T_f) = \sum_{(i,j) \in E_f} \|p_i^r - p_j^r - (p_i^h - p_j^h)\|^2, \quad (12)$$

where E_f denotes the edges of the tetrahedron, and p^r, p^h are vertex positions in the robot and human demonstration, respectively.

To distinguish cases where geometric structure is preserved but manipulation semantics differ, a relative-pose invariant is introduced. Using a local coordinate frame based on hand keypoints, the relative position of the object contact point is constrained:

$$L_{\text{pose}}(T_f) = \|R_r^\top (c_f^r - p_0^r) - R_h^\top (c_f^h - p_0^h)\|^2, \quad (13)$$

where p_0 is the palm reference point, and $R_r, R_h \in SO(3)$ represent the local hand frames of the robot and human demonstration, respectively. This term removes global translation and rotation while preserving the contact direction and spatial distribution relative to the hand, effectively maintaining manipulation semantics.

Finally, hand motion retargeting is formulated as the following constrained optimization problem:

$$q^* = \min_q \sum_f w_f (L(T_f(q))), \quad (14)$$

$$s.t. q_{\min} \leq q \leq q_{\max}, \quad (15)$$

where $q_{\min}, q_{\max}$ are joint limits, and q^* denotes the optimized joint configuration achieving minimal combined edge and pose loss.

D. Synchronized Arm-Hand Execution

After obtaining the optimized arm and hand control commands, ObjRetarget employs a unified temporal scheduler to coordinate the execution of both arms and hands. The arm motions are carried out using the anthropomorphically retargeted commands produced by the arm-level optimization, ensuring physically plausible and smooth trajectories. Hand motions are executed according to a contact-aware strategy. When no contact is detected, defined as a hand-object interaction distance greater than the threshold δ , the system follows the initialized hand retargeting commands. The threshold δ is set to 0.05cm, chosen based on the precision of the object tracking system to reliably distinguish actual contact from free motion. Once contact

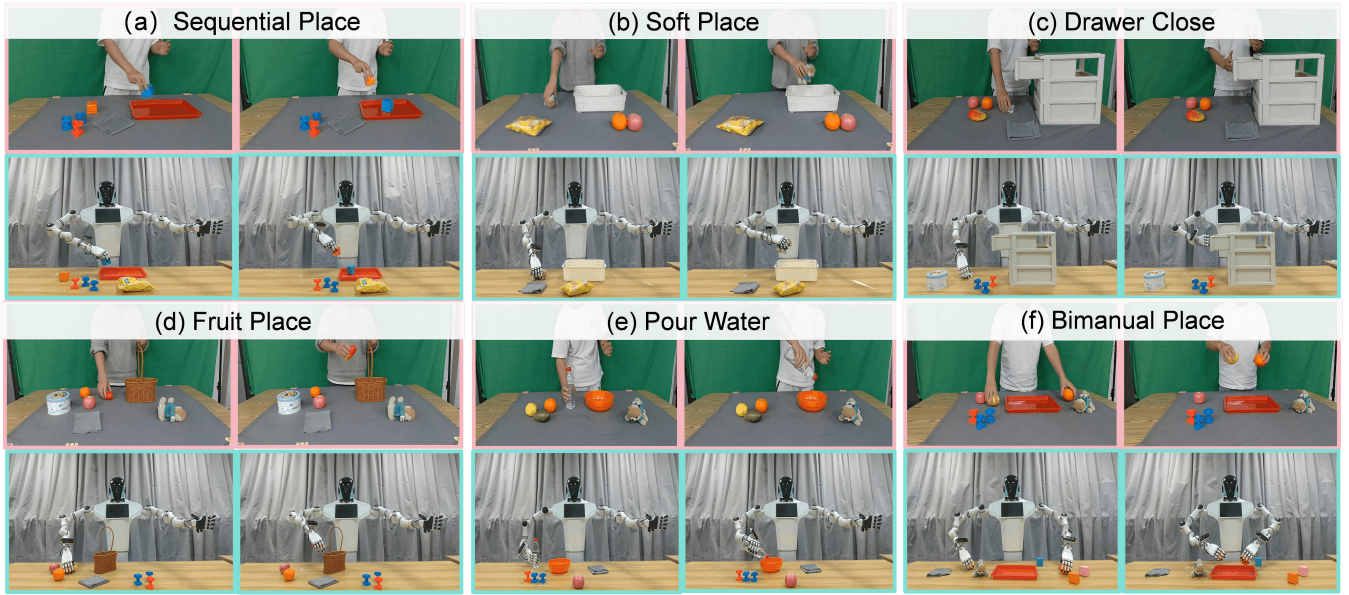


Fig. 4. Visualization of ObjRetarget on six real-world dexterous manipulation tasks. For each task, the top row shows the human demonstration and the bottom row shows the corresponding robot execution. From left to right: (a) Sequential Place: sequentially placing two cartons onto a tray, (b) Soft Place: grasping a plush toy and placing it into a basket, (c) Drawer Close: placing a medicine box into an open drawer and closing it, (d) Fruit Place: transferring a mango from the table to a fruit basket, (e) Pour Water: picking up a bottle, pouring water into a container, and returning it, and (f) Bimanual Place: bimanual grasping of an apple and a lemon followed by synchronized placement.

occurs, the hand control switches to the polyhedral-cluster-optimized commands to preserve local contact structures and manipulation semantics. After the contact is released, the system reverts to the initialized hand commands. This cycle repeats automatically for multiple contact events during an operation. By synchronizing arm and hand motions in this way, the scheduler ensures stable, natural, and generalizable execution of complex object-interaction tasks.

IV. EXPERIMENTS

A. Experimental Setup

To systematically evaluate the stability and generalization ability of ObjRetarget in real-world manipulation scenarios, we conduct experiments on a RealMan dual-arm robot platform across a range of everyday dexterous manipulation tasks. The tasks cover grasping, placement, sequential pouring control, and articulated-object interaction, enabling evaluation under varying interaction complexity and motion constraints. Specifically, the tasks include: a) sequentially placing two cartons onto a tray, b) grasping a plush toy and placing it into a basket, c) placing a medicine box into an open drawer and closing it, d) transferring a mango from the table to a fruit basket, e) picking up a bottle, pouring water into a container, and returning it, and f) bimanual grasping of an apple and a lemon followed by synchronized placement.

The experiments are performed on the RealMan dual-arm robot, where each arm has 6 DoF and is equipped with a 6-DoF Inspire dexterous hand. An Intel RealSense D435i depth camera is used for visual perception and experiment recording. The robot executes motions using joint-space position control for continuous trajectory tracking, ensuring smooth, synchronized, and stable dual-arm execution.

TABLE I
TASK SUCCESS RATE (%)

Task (/20f)	OKAMI [31]	ORION [32]	Ours
Soft Place	13	12	17
Pour Water	12	12	16
Fruit Place	14	13	16
Drawer Close	10	8	13
Sequential Place	12	9	14
Bimanual Place	13	8	15
Avg. Success Rate	61.6%	50.8%	75.8%

For evaluation, each task is repeated 20 times. In every trial, object poses are randomly initialized within the intersection of the camera field of view and the robot’s reachable workspace, and experiments are conducted in cluttered tabletop environments containing distractor objects. This setup allows us to assess robustness under realistic perception uncertainty.

We compare ObjRetarget with OKAMI [31] and ORION [32]. Since ORION was originally designed for parallel grippers, its generated end-effector trajectories cannot be directly applied to dexterous hands. To ensure fair comparison, we adapt its control interface by converting the gripper trajectory into a palm trajectory and incorporating hand pose and joint constraints in the inverse kinematics solver, enabling stable multi-finger grasping and placement under identical task settings.

All experiments were conducted using the PyTorch deep learning framework on a system equipped with an NVIDIA RTX 4090 GPU and an Intel(R) Core(TM) i7-11700KF CPU.

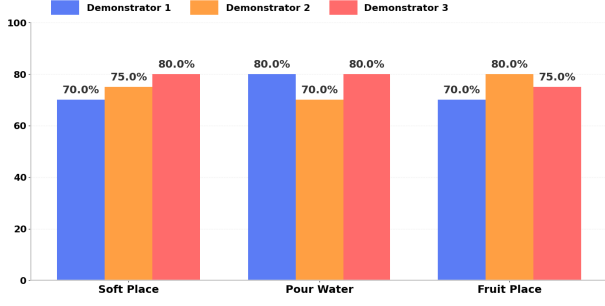


Fig. 5. Performance evaluation of ObjRetarget with videos from multiple demonstrators.

B. Comparative Experiments on Motion Retargeting

1) *Multi-method comparison*: We conduct a systematic comparative evaluation of ObjRetarget, OKAMI, and ORION across the full task set to assess their execution stability and task completion capability in complex dexterous manipulation scenarios. The visualization of ObjRetarget executions is shown in Fig. 4. In all experiments, object poses are randomly initialized within the workspace, requiring each method to autonomously perform the tasks under varying spatial configurations and visual observations. This setup allows us to realistically evaluate robustness and generalization performance in cluttered and uncertain environments.

As shown in Table I, ObjRetarget achieves the highest success rates across all tasks, with particularly clear gains in contact-sensitive and multi-stage scenarios such as Pour Water, Drawer Close, Sequential Place, and Bimanual Place. These results indicate that structured hand-object modeling improves contact stability, while unified arm-motion constraints enhance temporal consistency and coordination. OKAMI performs comparably on simpler placement tasks but degrades in operations requiring sustained contact or staged transitions. ORION yields the lowest overall performance, suggesting that without explicit motion and contact modeling, stable grasping and coordinated wrist behavior are difficult to maintain.

2) *Cross-demonstrator generalization*: To further evaluate the system’s robustness to individual differences, we tested ObjRetarget under cross-demonstrator conditions. Three demonstrators with distinct hand sizes, motion habits, and manipulation styles performed the same set of tasks. Without modifying the model structure or parameters, the demonstrations from all individuals were mapped onto the same robot platform for execution.

As shown in Fig. 5, ObjRetarget achieves stable performance across demonstrations from different individuals. Although variations in manipulation style and hand morphology introduce slight differences in task outcomes, the success rates remain consistently high for all demonstrators across the three tasks. In particular, each task maintains comparable performance ranges, indicating that the system does not rely on a specific motion style to function correctly.

C. Ablation Study

1) *Hand-object geometric consistency module*: We assess the contribution of structured contact modeling to task stability and success by comparing three configurations: removing the human-demonstration-based initialization (W/o Init Retarget), further removing hand-object geometric consistency (W/o Hand Geometry), and the full ObjRetarget. Evaluation metrics include the object slip distance [33], which measures the cumulative displacement of the object relative to the hand contact frame and thus reflects contact stability; geometric consistency [34], computed from normalized edge-length errors and relative pose errors of local hand-object contacts, which indicates how well the contact structure is preserved; and task success, the percentage of trials in which the task is completed successfully.

TABLE II
ABLATION STUDY OF HAND-OBJECT GEOMETRIC CONSISTENCY.

Method	Object Slip Distance(m) ↓	Geom. Consistency ↓	Task Success(%) ↑
W/o Init Retarget	0.025	0.183	65.0
W/o Hand Geom.	0.045	0.275	53.3
Ours	0.012	0.081	75.8

As shown in Table II, full ObjRetarget achieves the lowest geometric errors, minimal object sliding, and the highest success rates. Removing initialization slows convergence and reduces stability, while also removing geometric consistency causes significant slippage, contact failures, and posture collapse.

2) *Ablation of Arm Initialization and Anthropomorphic Constraints*: We evaluate the effects of the initialization-based retargeting and task-adaptive arm-plane regularization on motion accuracy and naturalness. Three configurations are compared: W/o Init Retarget (removing the initialization retargeting prior), W/o Arm-Plane Opt (removing arm-plane regularization), and full ObjRetarget. Performance is evaluated using four metrics: mean per-joint position error (MPJPE) for spatial accuracy [35], end-effector quaternion distance (Quat) for orientation accuracy [36], Fréchet distance for overall trajectory shape similarity [37], and task success rate for execution reliability.

TABLE III
ABLATION OF ARM INITIALIZATION AND ANTHROPOMORPHIC CONSTRAINTS

Method	MPJPE (m)↓	Quat (rad)↓	Fréchet (m)↓	Task Success (%)↑
W/o Init Retarget	0.125	0.275	0.313	60.8
W/o Arm-Plane Opt	0.103	0.337	0.251	57.5
Ours	0.088	0.192	0.124	75.8

As shown in Table III, full ObjRetarget achieves the best results across all metrics, producing trajectories closely aligned with human demonstrations. Removing initialization increases MPJPE and Fréchet distance due to a larger search

space, while removing the arm-plane regularization causes unnatural elbow motion and oscillations, reducing trajectory consistency.

V. CONCLUSION

This paper presents ObjRetarget, a human-to-robot re-targeting framework that decouples arm-level global motion from contact-sensitive hand manipulation by integrating polyhedral hand modeling and anthropomorphic arm constraints to achieve stable, natural, and generalizable dexterous manipulation. The framework is validated on the RMC-DA dual-arm robotic platform across diverse fine manipulation tasks, including grasping, pouring, object transfer, drawer closing, sequential placement, and bimanual coordination. Experimental results demonstrate that the proposed hand polyhedral cluster representation effectively preserves hand-object contact structures and manipulation semantics, while the reference-trajectory-guided arm optimization ensures accurate and anthropomorphically consistent global motion.

Despite these advantages, the current framework still has limitations: dynamic feedback in complex hand-object interactions is not yet fully exploited, and the method is presently limited to upper-limb manipulation. Future work will focus on integrating adaptive force feedback and extending ObjRetarget to whole-body robotic manipulation, aiming to enable more general and compliant dexterous manipulation systems.

REFERENCES

- [1] H. Xiong, Q. Li, Y. C. Chen, et al., "Learning by watching: Physical imitation of manipulation skills from human videos," in Proc. IEEE/RSJ Int. Conf. Intell. Robot Syst. (IROS), 2021, pp. 7827–7834.
- [2] Y. Qin, Y. H. Wu, S. Liu, et al., "DexMV: Imitation learning for dexterous manipulation from human videos," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2022, pp. 570–587.
- [3] Y. Liu, W. C. Shin, Y. Han, et al., "ImMimic: Cross-domain imitation from human videos via mapping and interpolation," arXiv preprint arXiv:2509.10952, 2025.
- [4] L. S. Li, J. Jiang, P. Ruppel, et al., "A mobile robot hand-arm teleoperation system by vision and IMU," in Proc. IEEE/RSJ Int. Conf. Intell. Robot Syst. (IROS), 2020, pp. 10900–10906.
- [5] S. Li, N. Hendrich, H. Liang, et al., "A dexterous hand-arm teleoperation system based on hand pose estimation and active vision," IEEE Trans. Cybern., vol. 54, no. 3, pp. 1417–1428, 2022.
- [6] A. Sivakumar, K. Shaw, and D. Pathak, "Robotic telekinesis: Learning a robotic hand imitator by watching humans on YouTube," arXiv preprint arXiv:2202.10448, 2022.
- [7] D. Antotsiou, G. Garcia-Hernando, and T.-K. Kim, "Task-oriented hand motion retargeting for dexterous manipulation imitation," in Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW), 2018.
- [8] M. Zhang, Q. Gao, Y. Lai, et al., "HR-GCN: 2D–3D whole-body pose estimation with high-resolution graph convolutional network from a monocular camera," IEEE Sens. J., 2025.
- [9] L. Yang, X. Huang, Z. Wu, et al., "OmniRetarget: Interaction-preserving data generation for humanoid whole-body locomanipulation and scene interaction," arXiv preprint arXiv:2509.26633, 2025.
- [10] H. Weng, Y. Li, N. Sobanbabu, et al., "HDML: Learning interactive humanoid whole-body control from human videos," arXiv preprint arXiv:2509.16757, 2025.
- [11] O. A. Andrychowicz, B. Baker, M. Chociej, et al., "Learning dexterous in-hand manipulation," Int. J. Robot. Res., vol. 39, no. 1, pp. 3–20, 2020.
- [12] A. Gupta, C. Eppner, S. Levine, et al., "Learning dexterous manipulation for a soft robotic hand from human demonstrations," in Proc. IEEE/RSJ Int. Conf. Intell. Robot Syst. (IROS), 2016, pp. 3786–3793.
- [13] W. Yang and W. Jin, "ContactSDF: Signed distance functions as multi-contact models for dexterous manipulation," IEEE Robot. Autom. Lett., 2025.
- [14] A. Rajeswaran, K. Lowrey, E. V. Todorov, et al., "Towards generalization and simplicity in continuous control," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017.
- [15] S. Calinon, "A tutorial on task-parameterized movement learning and retrieval," Intell. Serv. Robot., vol. 9, no. 1, pp. 1–29, 2016.
- [16] Z. Si, Z. Zhu, A. Agarwal, S. Anderson and W. Yuan, "Grasp stability prediction with sim-to-real transfer from tactile sensing," in Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS), 2022, pp. 7809–7816.
- [17] Y. Liang, et al., "Dynamic movement primitive based motion retargeting for dual-arm sign language motions," in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), 2021, pp. 8195–8201.
- [18] X. Xing, K. Maqsood, C. Zeng, et al., "Dynamic motion primitives-based trajectory learning for physical human-robot interaction force control," IEEE Trans. Ind. Inform., vol. 20, no. 2, pp. 1675–1686, 2023.
- [19] T. He, Z. Luo, W. Xiao, et al., "Learning human-to-humanoid real-time whole-body teleoperation," in Proc. IEEE/RSJ Int. Conf. Intell. Robot Syst. (IROS), 2024, pp. 8944–8951.
- [20] X. B. Peng, et al., "Learning agile robotic locomotion skills by imitating animals," arXiv preprint arXiv:2004.00784, 2020.
- [21] S. Li, et al., "Vision-based teleoperation of Shadow dexterous hand using end-to-end deep neural network," in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), 2019, pp. 416–422.
- [22] H. Zhang, et al., "Kinematic motion retargeting via neural latent optimization for learning sign language," IEEE Robot. Autom. Lett., vol. 7, no. 2, pp. 4582–4589, Apr. 2022.
- [23] B. Zheng, D. Liang, Q. Huang, et al., "Frame-by-frame motion retargeting with self-collision avoidance from diverse human demonstrations," IEEE Robot. Autom. Lett., 2024.
- [24] H. Zhang, et al., "Kinematic motion retargeting via neural latent optimization for learning sign language," IEEE Robot. Autom. Lett., vol. 7, no. 2, pp. 4582–4589, Apr. 2022.
- [25] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in Proc. Int. Conf. Artif. Intell. Stat. (AISTATS), 2011, pp. 627–635.
- [26] C. Finn, T. Yu, T. Zhang, et al., "One-shot visual imitation learning via meta-learning," in Proc. Conf. Robot Learn. (CoRL), 2017, pp. 357–368.
- [27] M. Shridhar, L. Manuelli, and D. Fox, "CLIPort: What and where pathways for robotic manipulation," in Proc. Conf. Robot Learn. (CoRL), 2022, pp. 894–906.
- [28] S. Zhu, R. Kaushik, S. Kaski, et al., "Imitation-guided multimodal policy generation from behaviourally diverse demonstrations," in Proc. IEEE/RSJ Int. Conf. Intell. Robot Syst. (IROS), 2023, pp. 1675–1682.
- [29] V. Ye, G. Pavlakos, J. Malik, and A. Kanazawa, "Decoupling human and camera motion from videos in the wild," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023.
- [30] Anonymous, "DexTele: A dual-arm dexterous teleoperation system based on motion retargeting and adaptive force control," in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), 2026, to appear.
- [31] J. Li, Y. Zhu, Y. Xie, et al., "OKAMI: Teaching humanoid robots manipulation skills through single video imitation," arXiv preprint arXiv:2410.11792, 2024.
- [32] Y. Zhu, A. Lim, P. Stone, and Y. Zhu, "Vision-based manipulation from single human video with open-world object graphs," arXiv preprint arXiv:2405.20321, 2024.
- [33] G. Garcia-Hernando, S. Yuan, S. Baek, et al., "First-person hand action benchmark with RGB-D videos and 3D hand pose annotations," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 409–419.
- [34] S. Song, A. Zeng, J. Lee, et al., "Grasping in the wild: Learning 6-DoF closed-loop grasping from low-cost demonstrations," IEEE Robot. Autom. Lett., vol. 5, no. 3, pp. 4978–4985, 2020.
- [35] R. Villegas, J. Yang, D. Ceylan, et al., "Neural kinematic networks for unsupervised motion retargeting," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 8639–8648.
- [36] R. Greer, N. Deo, and M. Trivedi, "Trajectory prediction in autonomous driving with a lane heading auxiliary loss," IEEE Robot. Autom. Lett., vol. 6, no. 3, pp. 4907–4914, 2021.
- [37] H. Alt and M. Godau, "Computing the Fréchet distance between two polygonal curves," Int. J. Comput. Geom. Appl., vol. 5, no. 1–2, pp. 75–91, 1995.