Give Me That: A Lightweight Retrieval-Based System for Interactive Point-to-Grasp Manipulation

He-Yang Xu[1]*, Zi-Yao Lin[1]*, Mingqi Gao[1], Yu-Hao Wei[1] and Xiu-Shen Wei[1]†

[1]Southeast University

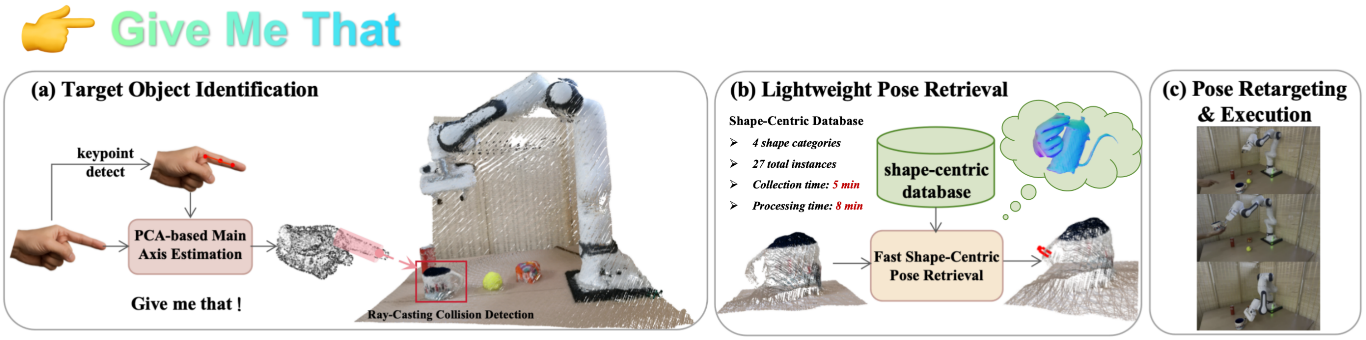
Give Me That
(a) Target Object Identification
keypoint detect
PCA-based Main Axis Estimation
Give me that !
Ray-Casting Collision Detection
(b) Lightweight Pose Retrieval
Shape-Centric Database
4 shape categories
27 total instances
Collection time: 5 min
Processing time: 8 min
shape-centric database
Fast Shape-Centric Pose Retrieval
(c) Pose Retargeting & Execution

Fig. 1: ***Give Me That*** is a lightweight retrieval-based point-to-grasp system with three stages: (a) Intuitive pointing enables explicit intent recognition; (b) fast pose retrieval from a shape-centric database; and (c) grasp retargeting and execution. By leveraging the shape-centric database, the system achieves fast and reliable human–robot interaction.

***Abstract*—Intuitive human-robot interaction (HRI) is crucial for enabling robots to act as assistive partners, yet existing manipulation systems remain limited. Traditional methods based on pose estimation or grasp sampling often lack stability and generalization, while recent end-to-end methods demand large annotated datasets and still generalize poorly. In this work, we present a lightweight system for point-to-grasp manipulation that builds a shape-centric grasp database within minutes from a few demonstrations, intuitively recognizes human intent through pointing gestures, and efficiently retrieves and retargets grasp poses for execution. Real-world experiments show high success rates on seen (100%) and unseen (92%) objects with inference latency as low as 0.148 s, demonstrating both effectiveness and real-time capability.**

I. Introduction

Intuitive human-robot interaction (HRI) is vital for enabling robots to serve as assistive partners, yet current manipulation systems remain limited. Pose estimation and grasp sampling often lack stability and generalization, while end-to-end learning demands large annotated datasets and still struggles with novel objects.

We propose a lightweight point-to-grasp system that builds a shape-centric database from few demonstrations, recognizes human intent through pointing gestures, and retrieves grasps efficiently for execution.

Specifically, our main contributions are:

- A shape-centric grasp database that captures grasp knowledge efficiently and is easily extensible.
- A lightweight retrieval-based framework for intuitive, safe, and controllable HRI.
- Real-world experiments showing high success on seen (100%) and unseen (92%) objects, with inference latency of 0.148 s.

II. Methods

A. Shape-Centric Database Construction

We construct a shape-centric grasp database enabling fast pose retrieval. It covers 4 shape categories (27 instances), built within 5 min collection and 8 min processing, and is easily extensible.

B. Retrieval-Based HRI Framework

We propose a lightweight retrieval-based HRI framework (Fig. 1), where the target is intuitively identified from pointing gestures, a suitable grasp is retrieved from the shape-centric database, and then retargeted for robot execution.

III. Results

Tab. I shows our method surpasses GraspNet on both seen and unseen objects, with success rates up to 100%. Tab. II further highlights its efficiency, achieving 0.148 s latency—faster than GraspNet (0.208 s) and far quicker than RAM (23.6 s).

TABLE I: Success Rates (%) on Seen and Unseen Objects.

Settings	**Objects**	**GraspNet**	**Ours**
Seen Objects	Mug	88%	**96%**
	Ball	84%	**100%**
	Sandbag	76%	**96%**
Unseen Objects	White mug	88%	**92%**
	Aluminum beverage can	80%	**100%**
	Cardboard box	80%	**88%**

TABLE II: Latency (s) Comparison of Different Methods.

	RAM	**GraspNet**	**Ours**
Inference Latency	23.578±0.059	0.208±0.012	**0.148**±0.001

Manuscript 5499 submitted to 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Received August 20, 2025.