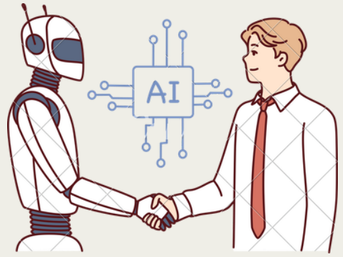
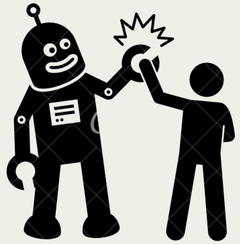
MapleGrasp: **Ma**sk-guided feature **P**ooling for **L**anguage-driven **E**fficient robotic **Grasp**ing

Submission to WACV 2026

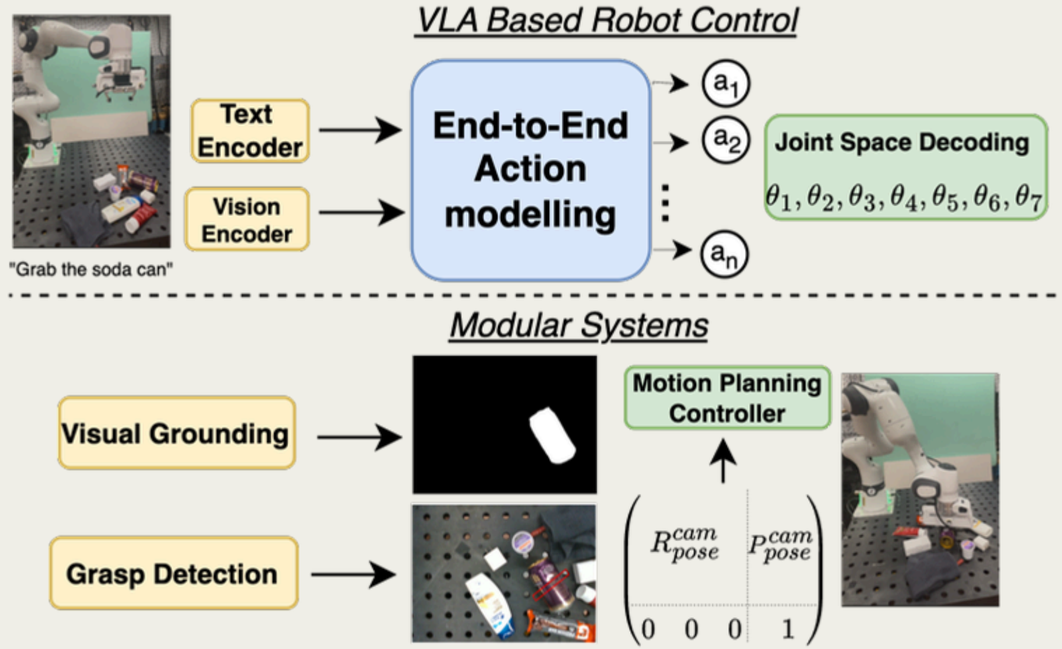
MOTIVATION



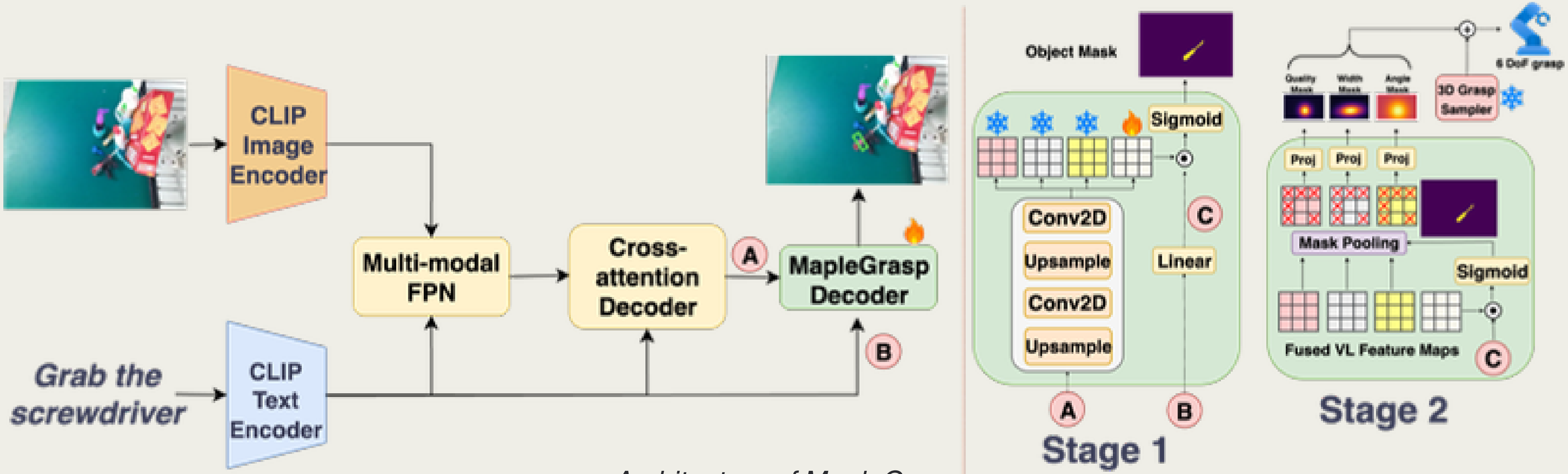
- Ability to associate natural language with real world objects is fundamental for Human-Robot interaction



- Human vocabulary is diverse; Visual grounding must be robust to variations in natural language queries



MODEL



Architecture of MapleGrasp

- **Stage 1:** Train the network to predict referred object masks from vision-language inputs, keeping the grasp prediction modules frozen.
- **Stage 2:** Fine-tune the entire network using the predicted masks to pool vision-language features for grasp quality, angle, and width prediction.

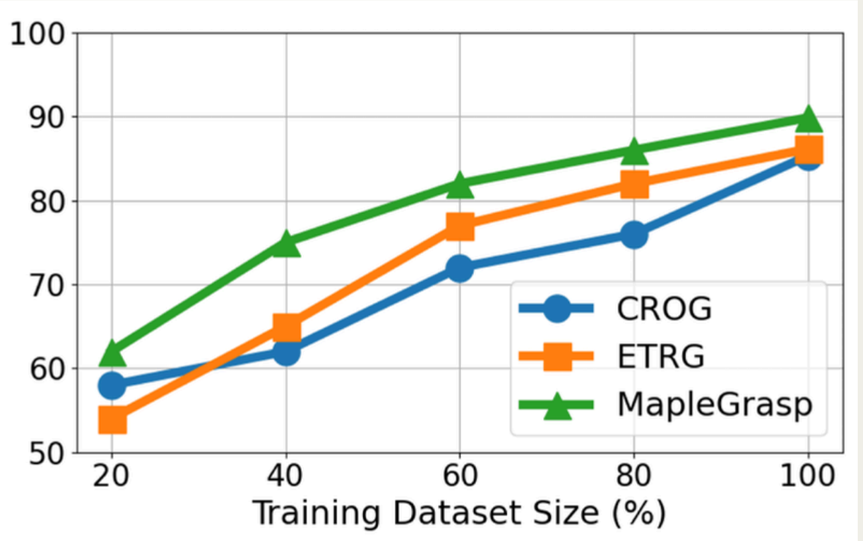
DATASETS

- **OCID-VLG**
  - 1,700+ cluttered tabletop scenes with 31 unique objects; widely adopted for language-driven grasp detection.
- **RefGraspNet (Contributed Dataset)**
  - Extends GraspNet-1B with LLM-generated grasping instructions using object attributes like color, shape, and position.
  - Contains 200M+ free-form queries, enabling diverse, robust, and human-guided grasping at scale.

Samples from RefGraspNet. Only top-down rectangular grasps (pink) shown for brevity.

RESULTS ON IMAGE BENCHMARKS

Model	Epochs to Conv.	OCID-VLG	RefGraspNet	
			Test Seen	Test Unseen
DetSeg + CLIP [1]	22	28.12 / 39.21	40.19 / 41.56	27.18 / 27.20
GR-ConvNet + CLIP [20]	15	9.73 / 15.41	34.19 / 46.56	30.19 / 32.17
SSG + CLIP [51]	41	33.51 / 34.70	48.78 / 50.10	14.36 / 14.89
Molmo+SAM2+GraspNet	—	63.22 / 69.14	68.85 / 70.27	66.23 / 69.41
CROG [43]	50	77.22 / 87.71	85.32 / 86.49	70.81 / 71.99
HiFi-CS [2]	44	70.54 / 79.12	73.27 / 74.55	62.13 / 63.28
ETRG [54]	49	82.28 / 91.12	86.13/88.12	68.11 / 70.05
Ours: MapleGrasp	32	**88.15 / 92.98**	**89.86 / 91.67**	**76.92 / 77.67**



Model	Train-Dataset	OCID-VLG		RefGraspNet	
		Top-1	Top-5	Top-1	Top-5
CROG	OCID-VLG	77.2	87.7	41.8	42.9
	RefGraspNet	**68.2**	**73.8**	85.32	86.49
MapleGrasp	OCID-VLG	88.2	92.9	43.2	45.6
	RefGraspNet	**79.5**	**81.7**	89.2	89.7

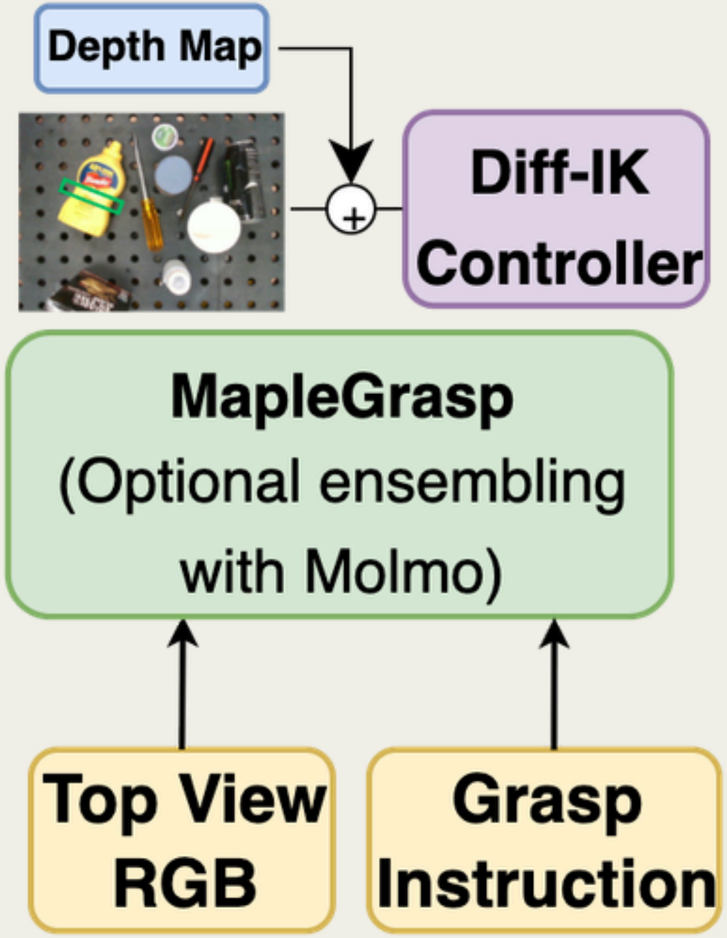
PHYSICS BASED SIMULATIONS

Model	SPATIAL	GOAL	OBJECT
Diffusion Policy [5]	77	69	**92**
Octo [40]	78	84	88
OpenVLA [21]	84	79	90
Otter [16]	84	82	89
MapleGrasp	**87**	**85**	90

Model	Train-Dataset	GOAL	OBJECT
OpenVLA	GOAL	79	0
	OBJECT	12	90
Otter	GOAL	82	22
	OBJECT	16	89
MapleGrasp	GOAL	85	**62**
	OBJECT	**68**	90

Task Instruction	OpenVLA	MapleGrasp
Grab the bbq sauce		
Pick up the alphabet soup		

REAL ROBOT EXPERIMENTS



(a) Seen Objects

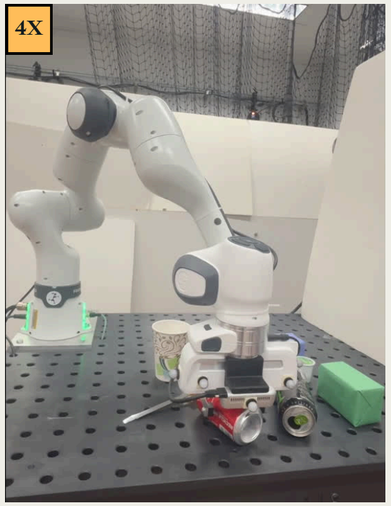
(b) Unseen Objects

ROLL-OUTS (1/4)

Grab the black soda can

Molmo+SAM+ GraspNet

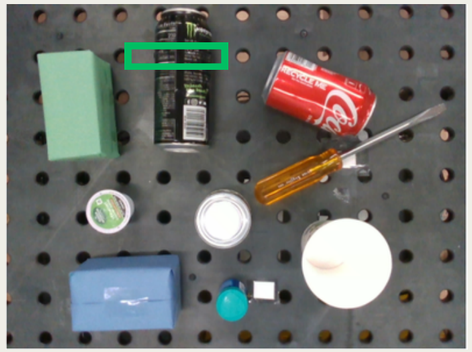
Incorrect grasp predicted for distractor (red can)



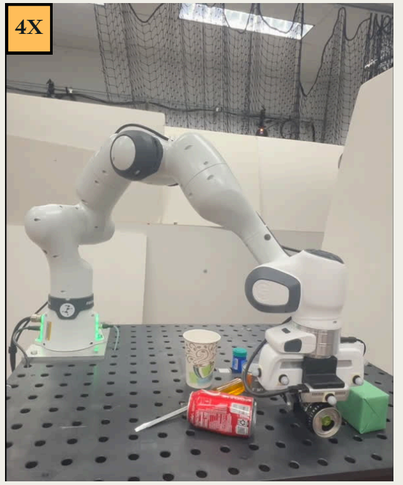
Failed Grasp

Grab the black soda can

MapleGrasp



Correct grasp for target object



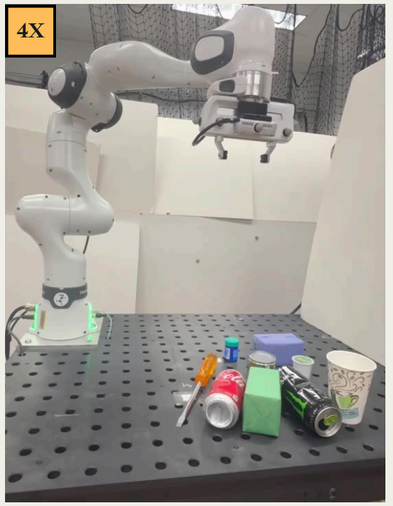
Successful Grasp

ROLL-OUTS (2/4)

Get me the yellow screwdriver

CROG

Unstable grasp predicted at boundary of object

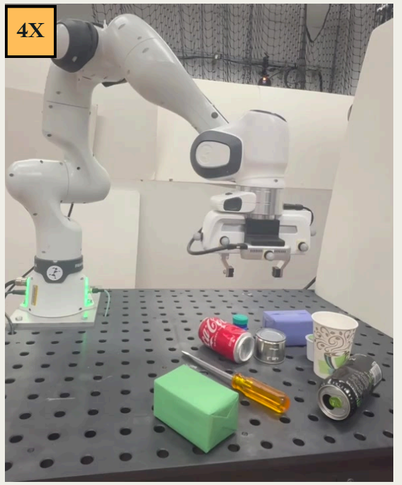


Failed Grasp

Get me the yellow screwdriver

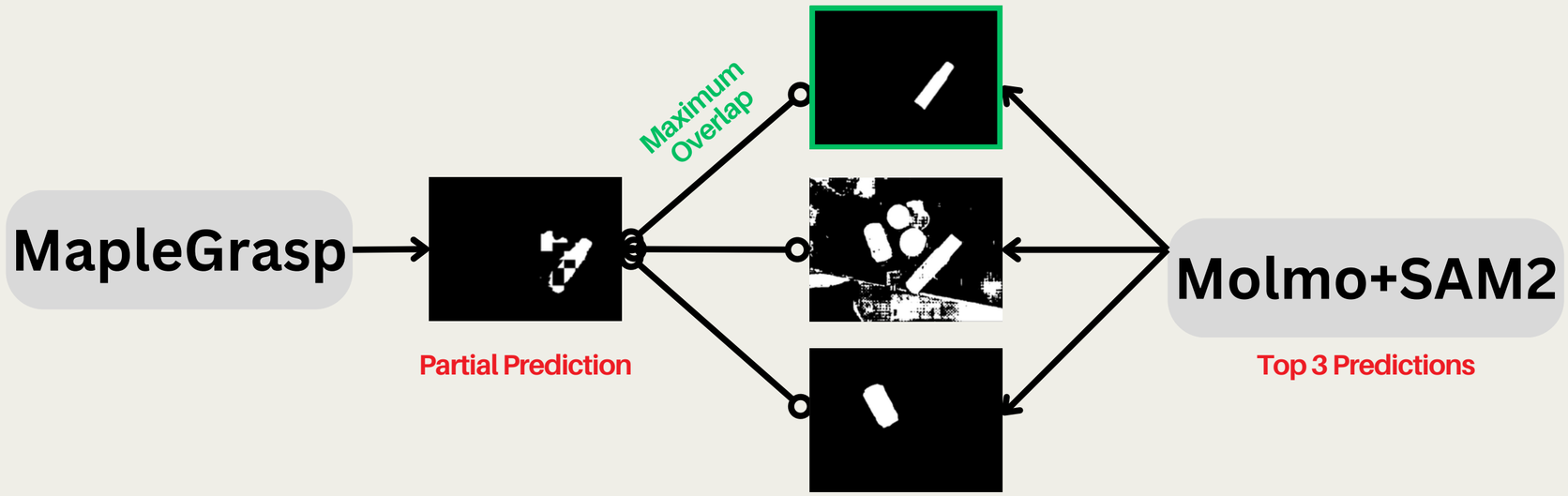
MapleGrasp

Correct grasp for target object near the handle



Successful Grasp

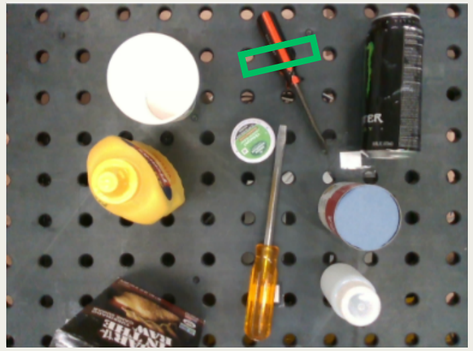
ROLL-OUTS (3/4)



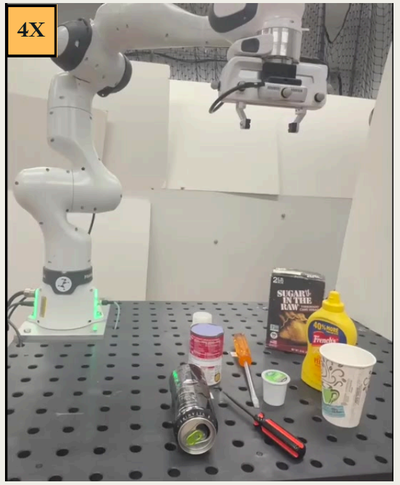
MapleGrasp+Molmo: Combining MapleGrasp with Molmo + SAM2 for improved performance with unseen objects

Get me the red and black screwdriver

MapleGrasp +Molmo



Identifies stable grasp for unseen object



Successful Grasp

ROLL-OUTS (4/4)

Pass me the yellow mustard

MapleGrasp

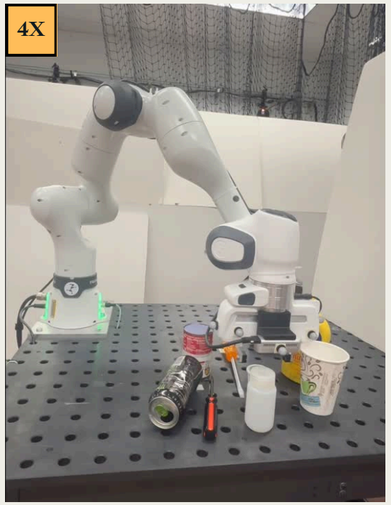
Model predicts correct grasp

Failed execution due to mis-aligned grasp pose

Grab the circular container to the left of the red screwdriver

MapleGrasp

Model incorrectly tries to grab the seen object next to the yellow screwdriver



Failed Grasp due to object detection failure !

CONCLUSION

- ***MapleGrasp*** *offers an effective and accurate approach for language-driven grasping using mask-guided feature pooling.*

- *Our approach matches the performance of billion-parameter VLAs while being data-efficient—trained on easily annotated* ***grasp poses instead of tele-operated demos.***

- *Validated in real-world, MapleGrasp demonstrates robustness to clutter, distractors, and* ***strong language grounding*** *for referred object grasping.*