

# UniRAP

## Towards Unified Part-level Physical Affordance Reasoning and Actionable Perception

Linfei Li<sup>1</sup> Ruining Hu<sup>1</sup> Lin Zhang<sup>1\*</sup> Zhong Wang<sup>1</sup>  
Fengyi Zhang<sup>4</sup> Ying Shen<sup>1</sup> Binqiang Wang<sup>2</sup> Xin Zhang<sup>3</sup>

<sup>1</sup> Tongji University <sup>2</sup> IEIT SYSTEMS (Beijing) Co., Ltd

<sup>3</sup> NAIS, Neusoft Corporation <sup>4</sup> The University of Queensland

# Language understanding must reach contact geometry

The instruction specifies a goal. The robot needs a part and a grasp.

## Actionable Perception for Real-World Robotic Manipulation

Q: Could you heat some water in the **teapot** and make me tea?  
A: Okay, I will grab the handle <|SEG|> of the teapot.



Q: Could you pour me some coffee using the **white cup**?  
A: Sure, I'll grab the handle <|SEG|> of the cup.



Q: Please help me find **a spoon**. I'd like to have some soup.  
A: Here's the spoon. I'm grabbing its handle <|SEG|>.



## Object-level grounding

Which object?

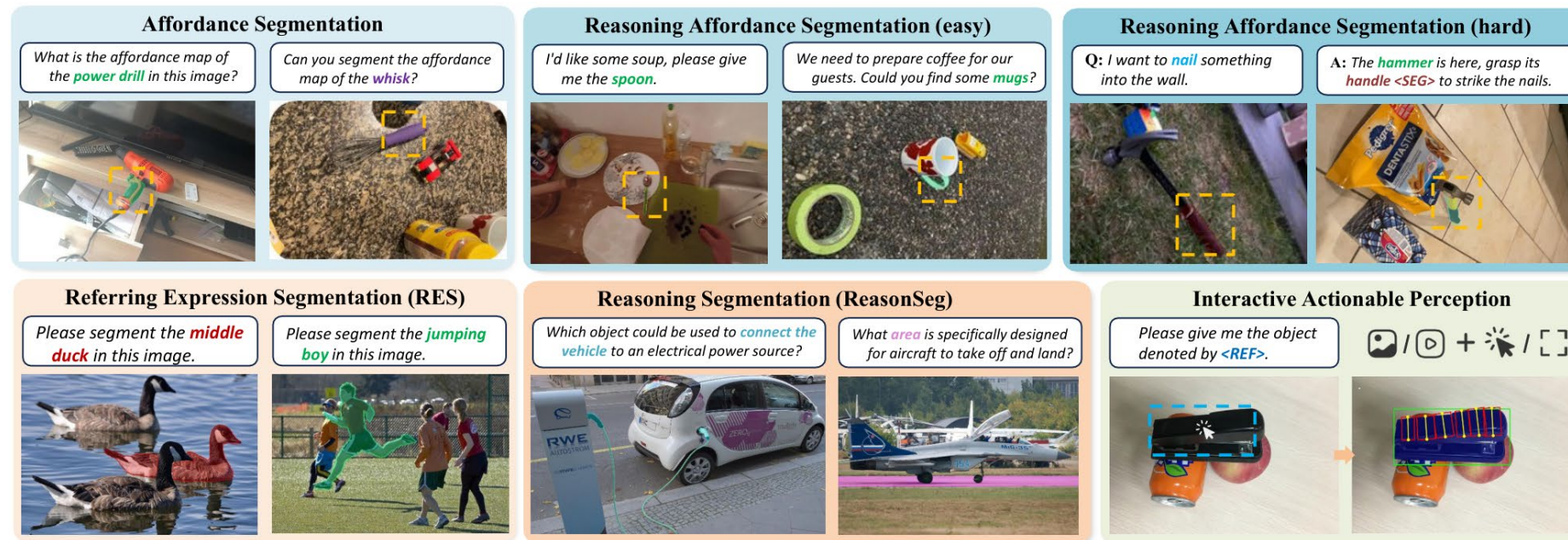
## Part-level affordance

Which functional region?

## Actionable perception

Where and how to contact?

# Task definition and scope



## Inputs

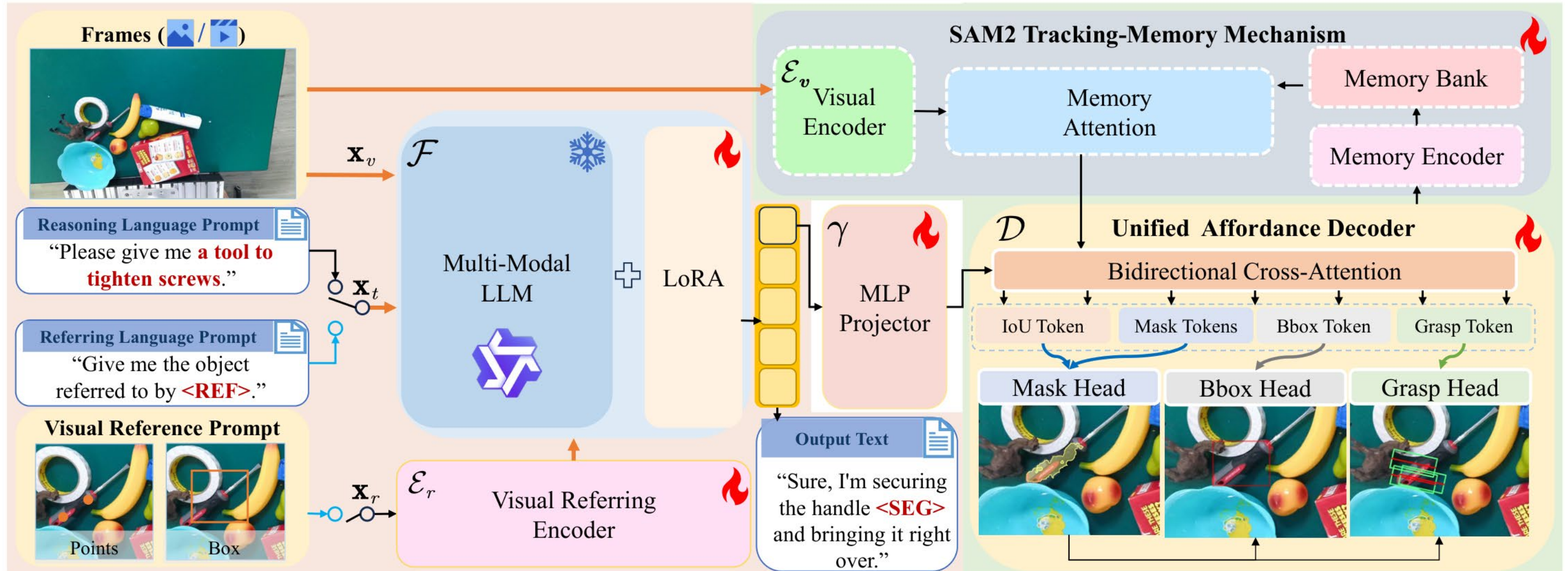
Image or video, language instruction, optional points or boxes.

## Outputs

Affordance mask, bounding box and 4-DoF grasp candidates.

A perception interface for downstream robotic execution.

# UniRAP architecture



Qwen2.5-VL (3B) + SAM2 + a shared interface for dense interaction prediction

# Text and visual prompts share the <SEG> interface

---

## Language referring

The instruction describes the target or its intended function.

## Visual referring

Points or boxes identify the target when language annotations are unavailable.

## Visual Prompt encoding

$$c = (x, y, t)$$

$$\phi(c) = [\sin(Bc), \cos(Bc)]$$

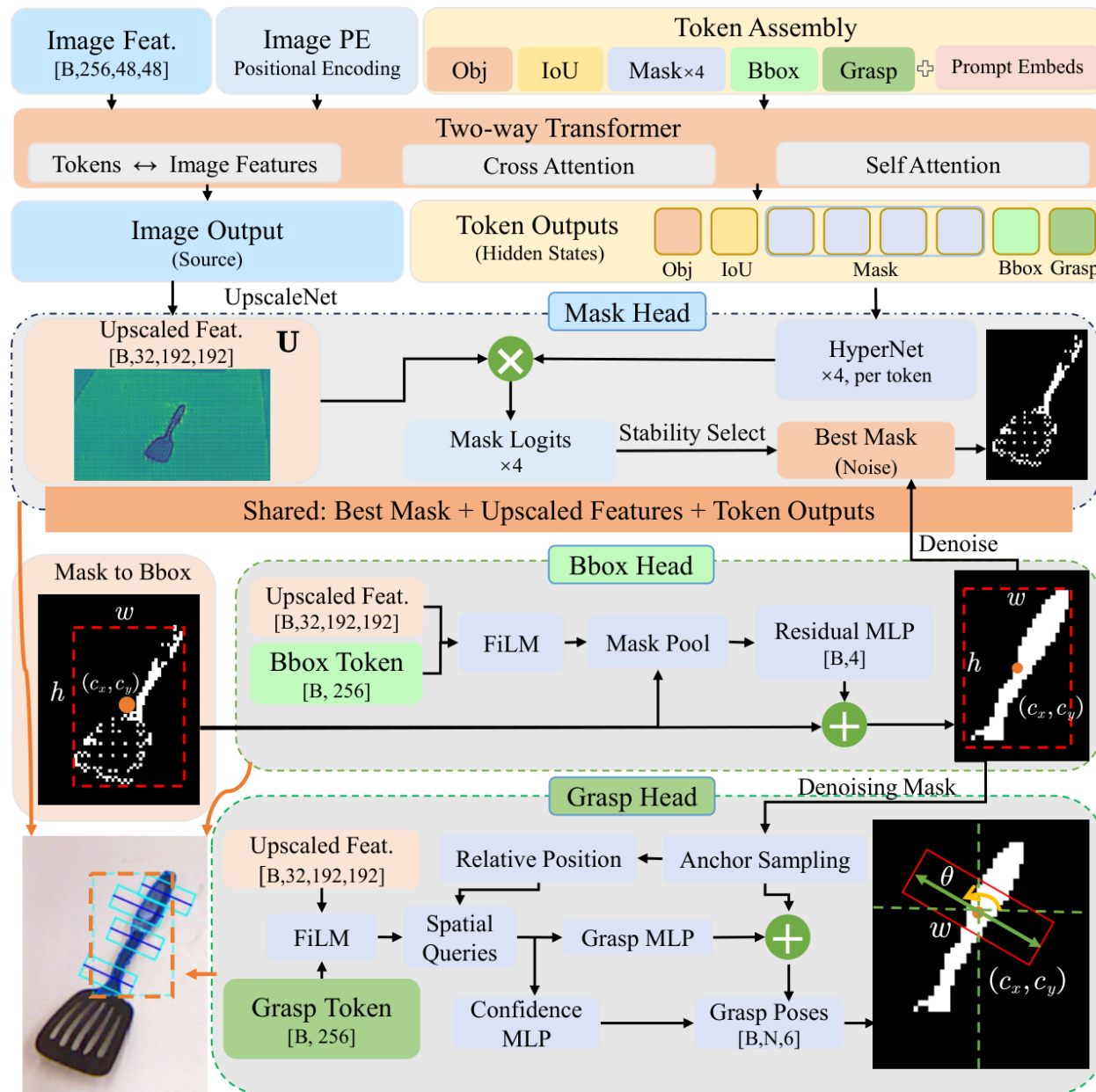
## Shared decoder prompt

$$h_{SEG} = \text{MLP}(\tilde{h}_{SEG})$$

$$(\hat{M}, \hat{b}, \hat{g}) = D(h_{SEG}, f_{mem})$$

**The shared token space enables transfer across different forms of supervision.**

# Unified Affordance Decoder



## Shared representations

Two-way attention exchanges information between task tokens and dense features.

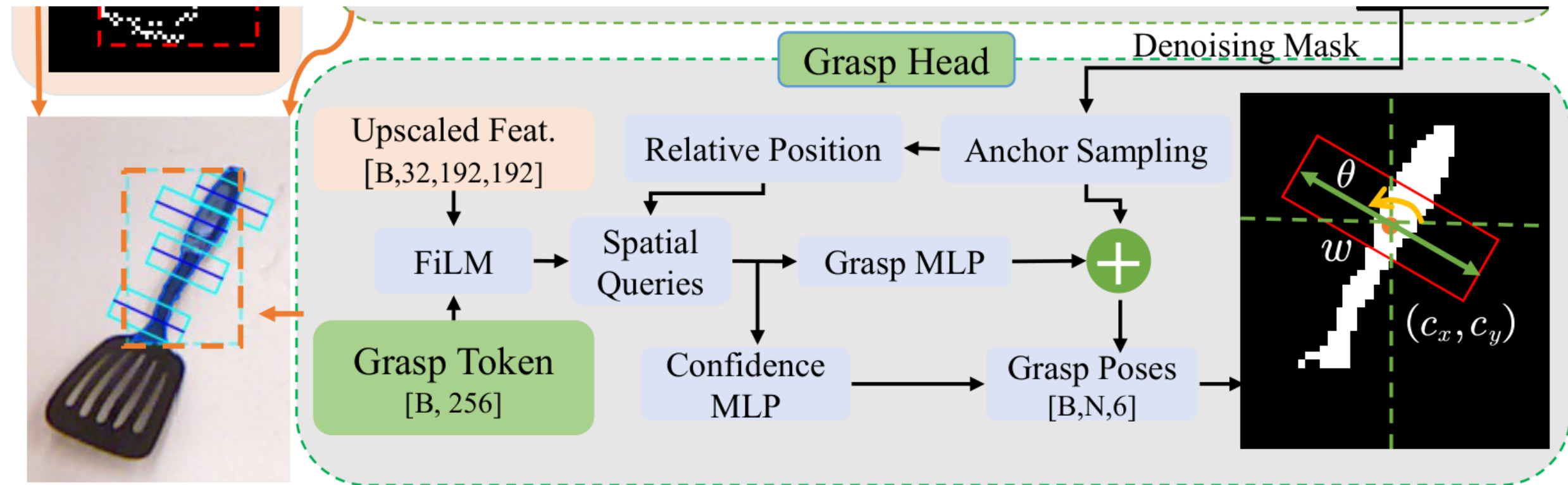
## Mask-guided localization

The mask initializes the box and supports residual box refinement.

## Contact prediction

Foreground features support direct grasp regression and confidence estimation.

# Grasp regression in the foreground region



## 32 foreground anchors

FPS sampling and relative position encoding

## Geometry regression

$(\Delta c_x, \Delta c_y, \cos 2\theta, \sin 2\theta, w)$   
Anchor + residual refinement

## Candidate confidence

A separate head estimates the quality score  $q$ .

# Curriculum transfer training



## Coarse to fine

Generic segmentation develops spatial features before contact prediction.

## Reduced text dependence

Visual referring prompts unlock additional grasp supervision.

**Final stage: optimize the Unified Affordance Decoder.**

# Training objectives and implementation

| Prediction   | Supervision                        |
|--------------|------------------------------------|
| Language     | Autoregressive cross-entropy       |
| Mask         | Focal + Dice, IoU and objectness   |
| Bounding box | L1 + generalized IoU               |
| Grasp        | Huber + confidence BCE             |
| Consistency  | Mask-in-box + grasp-center-in-mask |

## Geometry preservation

Center padding to  $768 \times 768$  avoids direct-resize distortion.

## Adaptation

LoRA rank 128,  $\alpha = 256$ .  
8 frames per video instance.

## Training hardware

**8 × A800 (80 GB)**

# Evaluation across perception and interaction

| Capability            | Benchmarks / setting             | Reported metrics                     |
|-----------------------|----------------------------------|--------------------------------------|
| Image segmentation    | RefCOCO / + / g, ReasonSeg       | cloU, gloU                           |
| Video segmentation    | ReVOS validation                 | J, F, J&F                            |
| Affordance grounding  | HANDAL, GraspNet, 3DOI           | gloU, cloU                           |
| Actionable perception | RealVLG: Seen / Similar / Novel  | Box IoU, Seg. F $\beta$ , grasp gAcc |
| Robotic grasping      | Franka FR3, 13 household objects | Physical grasp success (%)           |

Zero-shot affordance evaluation includes GraspNet Novel and 3DOI.

# General image segmentation remains competitive

RefCOCO / + / g: validation cloU. RefCOCOg: val(U). ReasonSeg: gloU.

| Method    | Size | RefCOCO | RefCOCO+ | RefCOCOg | ReasonSeg |
|-----------|------|---------|----------|----------|-----------|
| VideoLISA | 3.8B | 73.8    | 63.4     | 68.3     | 61.4      |
| Sa2VA     | 4B   | 78.9    | 71.7     | 74.1     | 52.3      |
| UniPixel  | 7B   | 80.8    | 75.3     | 76.4     | 60.5      |
| UniRAP    | 3B   | 80.9    | 75.1     | 76.1     | 57.9      |

Strong referring segmentation at 3B scale. ReasonSeg remains a limitation relative to VideoLISA.

# Video segmentation on ReVOS

J&F on ReVOS validation

| Method      | Size | Referring | Reasoning | Overall |
|-------------|------|-----------|-----------|---------|
| HyperSeg    | 3B   | 58.5      | 53.0      | 55.7    |
| InstructSeg | 3B   | 57.0      | 51.9      | 54.5    |
| UniPixel    | 7B   | 65.8      | 61.5      | 63.7    |
| UniRAP      | 3B   | 64.1      | 59.6      | 61.9    |

61.9 overall J&F

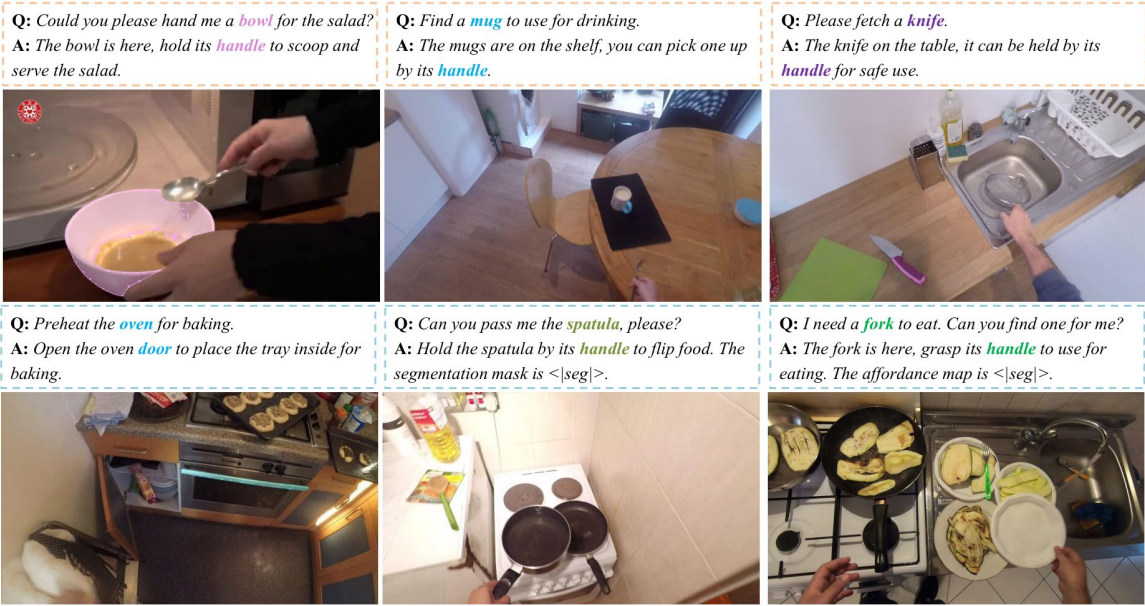


Competitive with larger models while extending to contact prediction.

# Affordance generalization on unseen objects

gloU, zero-shot evaluation

| Method      | GraspNet Novel | 3DOI referring | 3DOI reasoning |
|-------------|----------------|----------------|----------------|
| RAGNet (7B) | 45.6           | 37.4           | 38.1           |
| UniRAP (3B) | 51.6           | 52.1           | 55.7           |



+17.6

gloU points on 3DOI reasoning vs. RAGNet

# Actionable perception on RealVLG

## Grasp gAcc across object splits

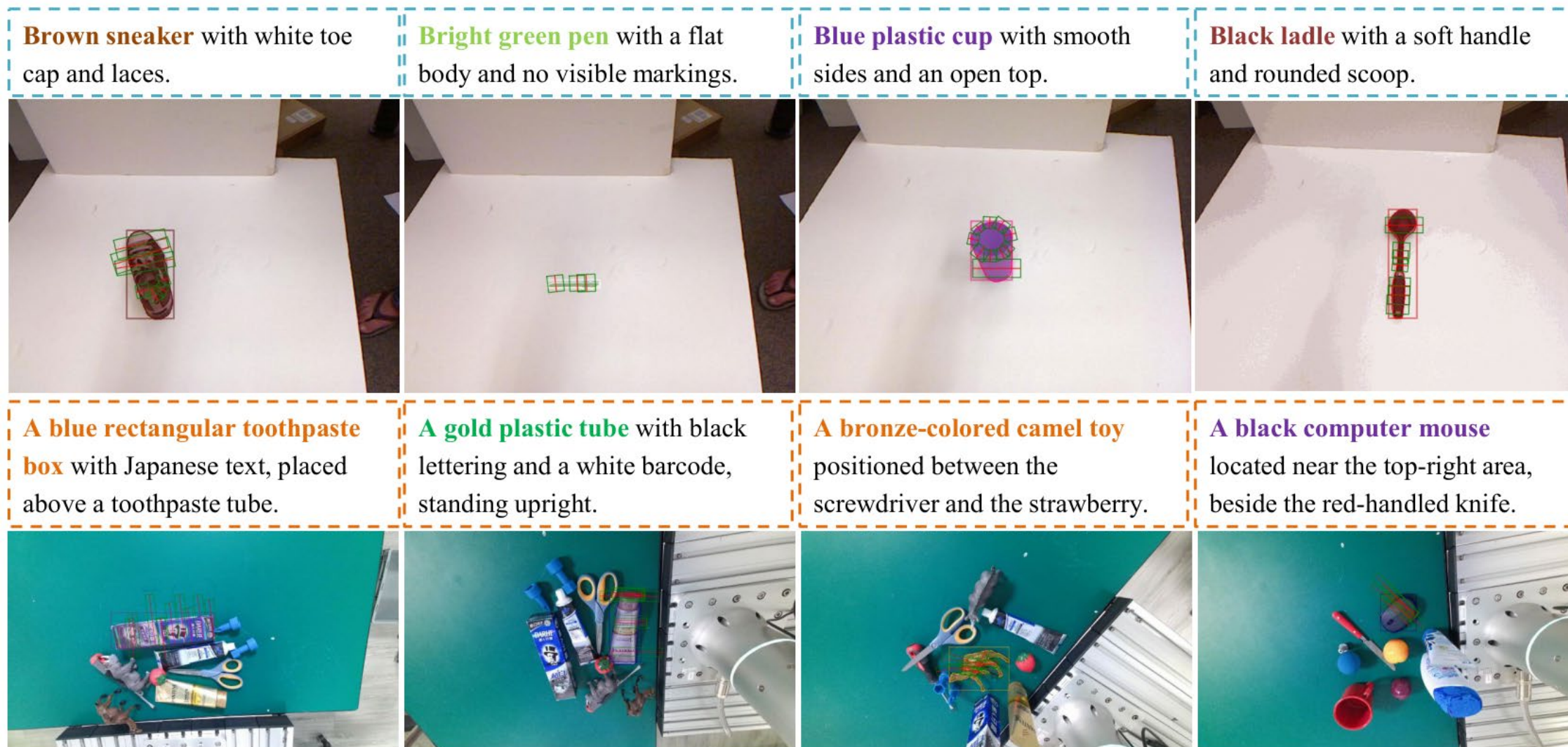
| Method               | Seen | Similar | Novel |
|----------------------|------|---------|-------|
| RealVLG-R1 GRPO (7B) | 39.1 | 36.2    | 18.9  |
| RealVLG-R1 GSPO (7B) | 32.8 | 30.2    | 18.3  |
| UniRAP (3B)          | 56.7 | 56.7    | 63.3  |

| Novel split          | Bbox gloU | Segmentation Fβ |
|----------------------|-----------|-----------------|
| RealVLG-R1 GSPO (7B) | 88.5      | 52.0            |
| UniRAP (3B)          | 68.4      | 75.8            |

+44.4

Novel grasp gAcc points  
vs. best RealVLG-R1

# Language-conditioned actionable predictions



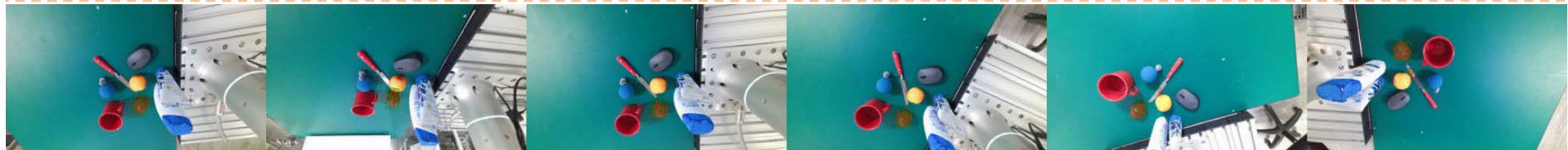
The same output interface supports explicit object descriptions and spatial referring expressions.

# Spatiotemporal actionable perception

**A gold plastic tube** with black lettering and a white barcode, labeled 'PANTENE', resting on the toothpaste box.



A small, round, dark purple object resembling **a plum**, located near the red mug and yellow ball.



**A small metallic component** with ridges and a cylindrical shape, resembling a bolt or screw, located above the peach-shaped object.



Memory attention combines current features with historical target information.

# What enables accurate grasp prediction?

| Variant                                   | Novel grasp gAcc |
|-------------------------------------------|------------------|
| Without visual referring pre-training     | 26.4             |
| Without actionable affordance fine-tuning | 19.6             |
| Without Bbox Head                         | 24.7             |
| Heatmap Grasp Head                        | 13.4             |
| Full UniRAP                               | 63.3             |

## Transfer learning

Visual referring links grasp supervision to the shared prompt space.

## Direct regression

The grasp head matters beyond mask quality.

# Real-world robotic grasping



Franka FR3 + RealSense D435i  
13 objects, 10 trials per object

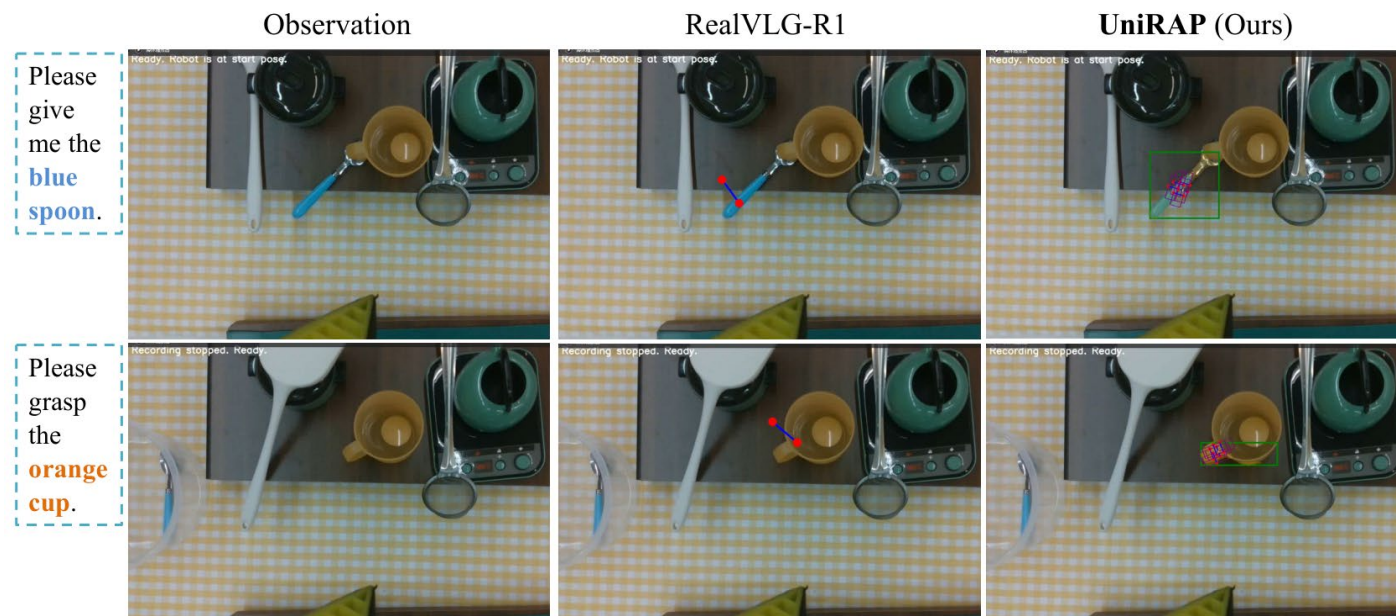
| Method             | Success (%) |
|--------------------|-------------|
| RAGNet + GraspNet  | 43.8        |
| UniRAP + GraspNet  | 55.4        |
| RealVLG-R1 (4-DoF) | 62.3        |
| UniRAP (4-DoF)     | 88.5        |

88.5%

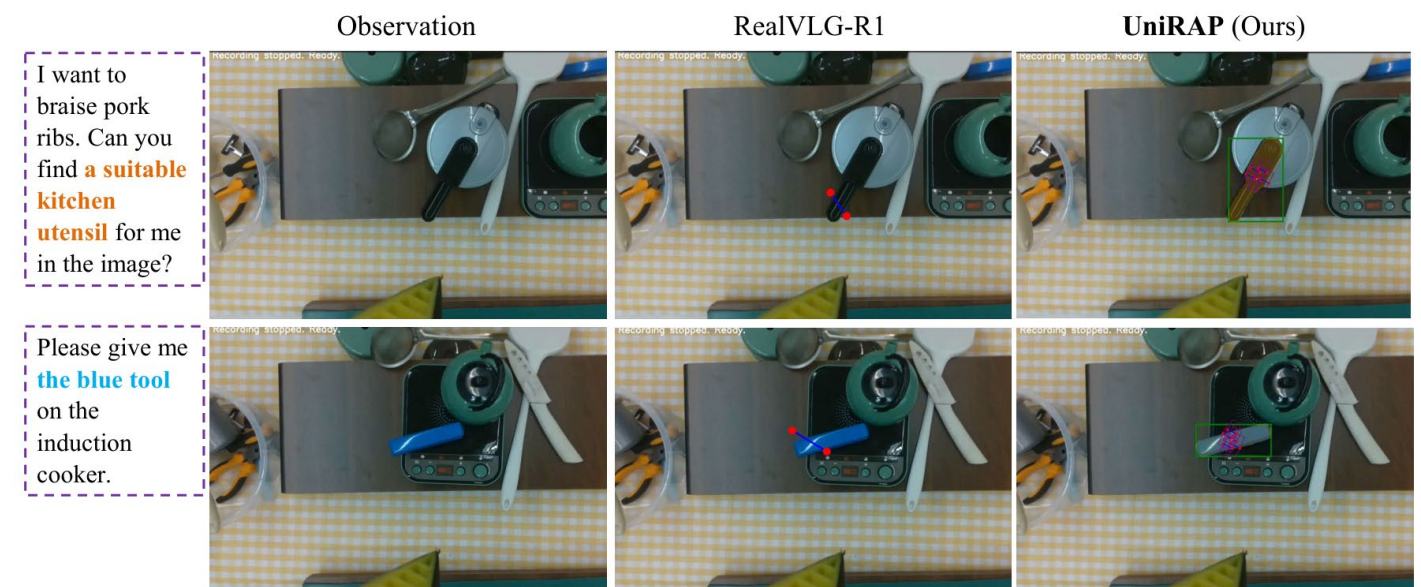
Average physical grasp success

# Referring and reasoning grasps on the robot

## Referring instructions



## Reasoning instructions



Language-conditioned contact geometry supports task-relevant grasp selection.

# Conclusions and future work

---

## Unified representation

A shared token interface connects language and visual prompts to dense interaction prediction.

63.3

Novel benchmark grasp gAcc

## Transfer across tasks

Segmentation features and visual referring pre-training support learning with limited language–grasp data.

88.5%

Real-world grasp success

## Current scope and next step

Extend 4-DoF actionable perception toward full 6-DoF orientation and dexterous interaction.

[lif314.github.io/projects/unirap](https://lif314.github.io/projects/unirap)