ActionVOS: Actions as Prompts for Video Object Segmentation

UTokyo
東京大学 生産技術研究所 Institute of Industrial Science, The University of Tokyo

Liangyang Ouyang, Ruicong Liu, Yifei Huang, Ryosuke Furuta, Yoichi Sato

Institute of Industrial Science, The University of Tokyo

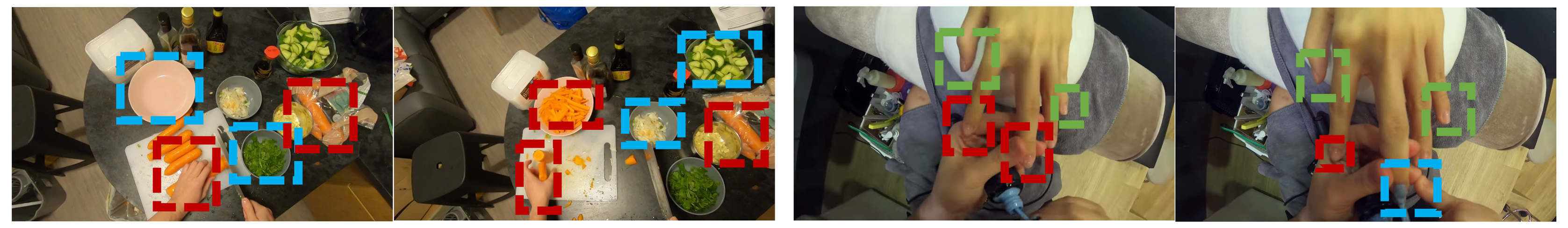
Github page   Follow us!

ECCV EUROPEAN CONFERENCE ON COMPUTER VISION MILANO 2024

Introduction

Referring video object segmentation (**RVOS**) aims at segmenting target objects using natural language expressions.

Challenges: Existing RVOS benchmarks primarily rely on **static attributes** such as object names and colors to describe the target objects.
In complex scenarios where **redundant instances** coexist or **object state changing**, such static attributes can not identify the target objects.



Static attributes: carrot bowl   nail pink nail blue nail

Key Idea: Human actions precisely describe the active objects.

Human actions: "put carrot in bowl"   "paint nail"

Our Solution: This work propose a novel action-aware RVOS setting, **ActionVOS**, segmenting **only active objects** by adding **human actions** as language prompts.

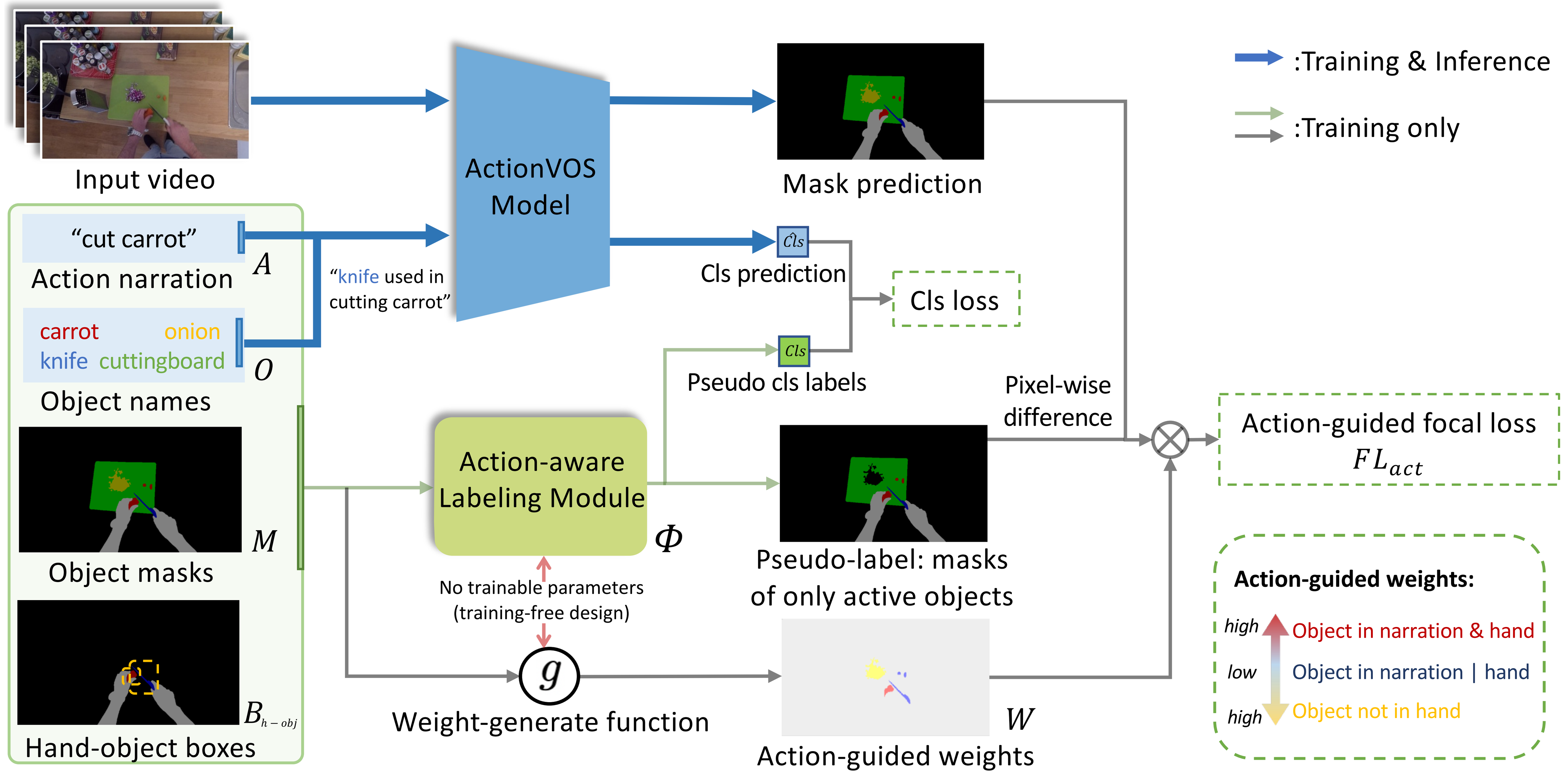
ActionVOS Problem Setting

Input:
- Video clip.
- Arbitrary object names.
- Action prompt describing the human action.

Output:
Masks of only active objects corresponding to the action prompt.

Definition of active objects:
- Objects described by the action prompt.
- Hands and hand-tools used in the action.
- Containers and contents interacted in the action.

Proposed Method for ActionVOS



Key Challenge: training an ActionVOS model **with existing readily-available annotations.** (A, O, M, B_{h-obj})

ActionVOS Model: Any RVOS model with an additional classification head.

Action-aware Labeling Module Φ:

Generate pseudo-labels of positive/negative objects.

$$Cls(O_i) = \begin{cases} 1, O_i \in A \\ 1, M(O_i) \cap B_{h-obj} \neq \emptyset \\ 0, otherwise \end{cases}$$

Pseudo positive label: 1) object whose name mentioned in the action prompt, e.g., carrot
2) object whose mask intersect with hand-object bounding boxes, e.g., knife, cutting board

Action-guided Focal Loss FL_{act}:

Modified segmentation focal loss by adjusting the pixel-wise weights W.
It is designed to reduce the impact of false positives in pseudo-labels. E.g., W(carrot in hand) > W(carrot on board).

ActionVOS Quantitative Results

ActionVOS results on VISOR. * serve as the upper bound of p-mIoU.

Model	Setting	ActionPrompt	p-mIoU↑	n-mIoU↓	gIoU↑	Acc↑
RF-R101	RVOS*	✗	67.7	54.2	43.8	59.1
	ActionVOS	✗	56.3	19.9	66.8	72.9
	ActionVOS	✓	**65.4**	**19.0**	**70.9**	**82.4**
RF-SwinL	RVOS*	✗	71.8	59.7	46.8	59.4
	ActionVOS	✗	64.4	28.2	65.1	72.8
	ActionVOS	✓	**69.1**	**24.6**	**70.3**	**80.7**
RF-VSwinB	RVOS*	✗	70.5	58.5	45.6	59.2
	ActionVOS	✗	61.6	25.2	65.7	72.5
	ActionVOS	✓	**68.2**	**22.0**	**70.6**	**81.2**

+5-10% p-mIoU gIoU with action prompts

-34% n-mIoU comparing to RVOS (less mis-segmentation on inactive objects)

ActionVOS results on VOST and VSCOS.

Model Dataset	Setting	AP	p-mIoU↑
RF-R101 VOST	RVOS	✗	29.3
	ActionVOS	✗	9.0
	ActionVOS	✓	**32.3**
RF-R101 VSCOS	RVOS	✗	46.4
	ActionVOS	✗	22.5
	ActionVOS	✓	**49.4**

Higher mIoU for state-changed objects

ActionVOS Quantitative Results

Comparison with baseline methods:

Method	p-mIoU↑	n-mIoU↓	gIoU↑	Acc↑	VOST		VSCOS	
					p-mIoU	p-cIoU	p-mIoU	p-cIoU
HOS	56.2	**11.4**	68.8	77.0	19.4	13.1	34.4	24.1
RVOS+Φ	65.3	35.2	60.4	75.1	29.3	17.5	46.4	44.9
Ours	**65.4**	19.0	**70.9**	**82.4**	**32.3**	**22.8**	**49.4**	**49.6**

+10% p-mIoU HOS
-16% n-mIoU RVOS+Φ

Evaluations on unseen actions:

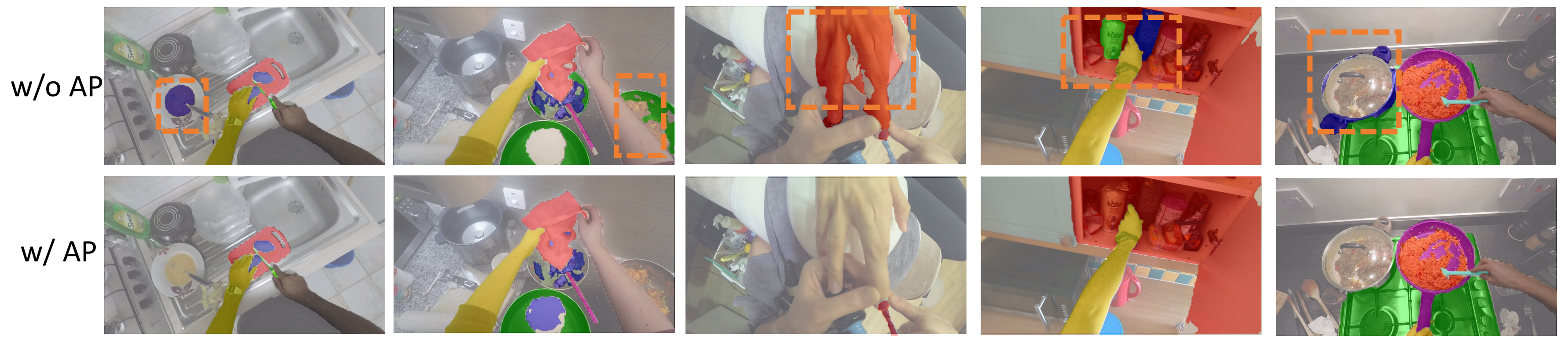
Method	p-mIoU↑	n-mIoU↓	gIoU↑	Acc↑	VOST		VSCOS	
					p-mIoU	p-cIoU	p-mIoU	p-cIoU
HOS	51.9	**9.0**	64.9	72.0	13.6	11.4	42.7	38.8
RVOS	60.0	49.0	42.9	65.3	18.6	12.6	31.5	21.4
Ours	**60.3**	21.0	**66.1**	**79.7**	**22.5**	**18.0**	**44.9**	**43.1**

+20% gIoU on unseen actions

+3-12% mIoU on unseen state changes

ActionVOS Qualitative Results

ActionVOS results trained w/ and w/o action prompts.   mis-segmentation of inactive objects



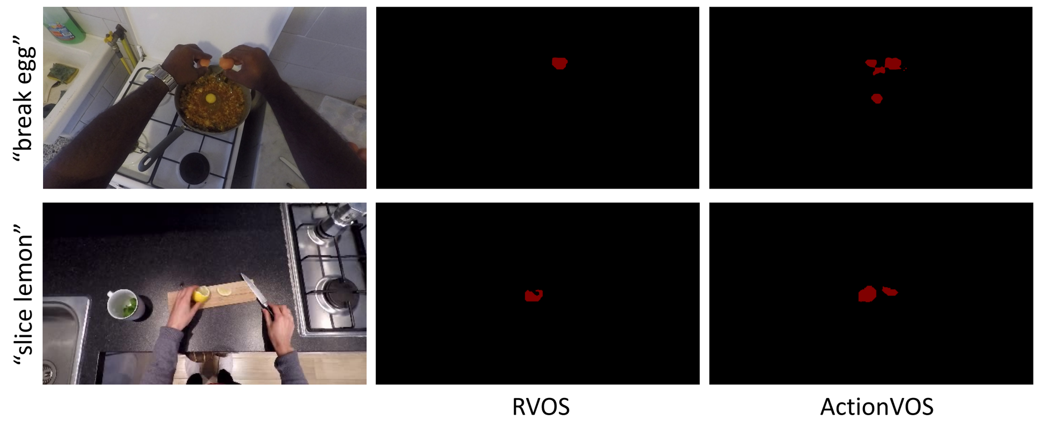
"cut potato"   "put paneer in pan"   "paint nail"   "open cupboard"   "mix food"

ActionVOS for different actions in same scene.

Input object names:
knife pizza olive
left hand right hand
chopping board jar

"pick up knife"   "put down jar"   "chop olives"   "put olives on pizza"

ActionVOS results on state-changed objects.



ActionVOS video visualization