CVPR
JUNE 17-21, 2024
SEATTLE, WA

Learning to Segment Referred Objects from Narrated Egocentric Videos

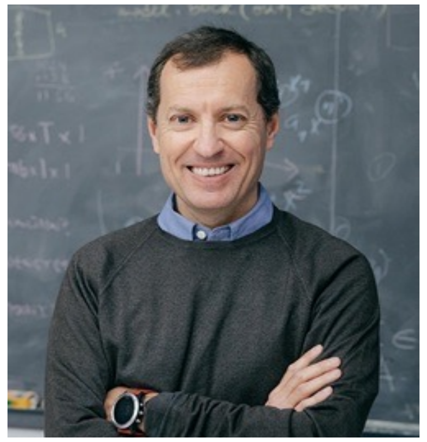

Yuhan Shen*


Huiyu Wang


Xitong Yang


Matt Feiszli


Ehsan Elhamifar*


Lorenzo Torresani


Effrosyni Mavroudi

FAIR, Meta
*Northeastern University

Poster: Thursday, 17:15 - 18:45, #460

Narration-based Video Object Segmentation

Input:

Egocentric Video Clip w/ Narration

Output:

Segmentation per Object Phrase

"C uses a **spoon** to spread the **minced meat** on the **cheese mixture** in a **lasagna dish**"

NVOS task evaluates the ability to ground referred objects in egocentric video at pixel-level

Example application: AI assistant highlighting required objects for a recipe step

Video Source: Epic-Kitchens

Data Source: VISOR-NVOS (see slide 9)

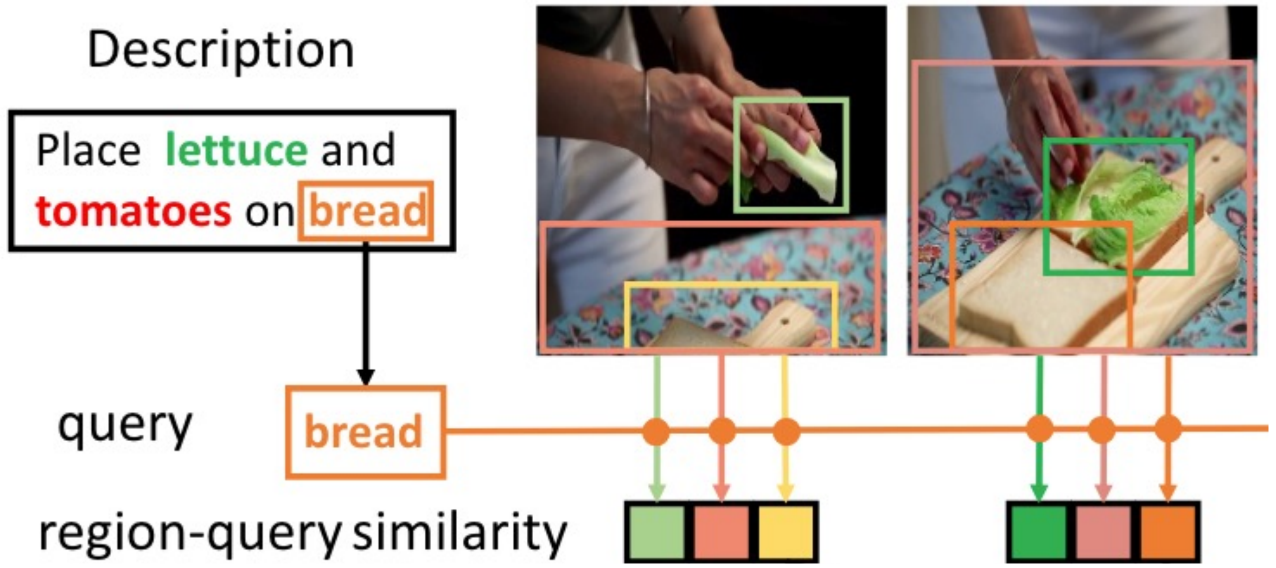
Related Work

Bounding-box-level grounding

spatial supervision

RefEgo [1]

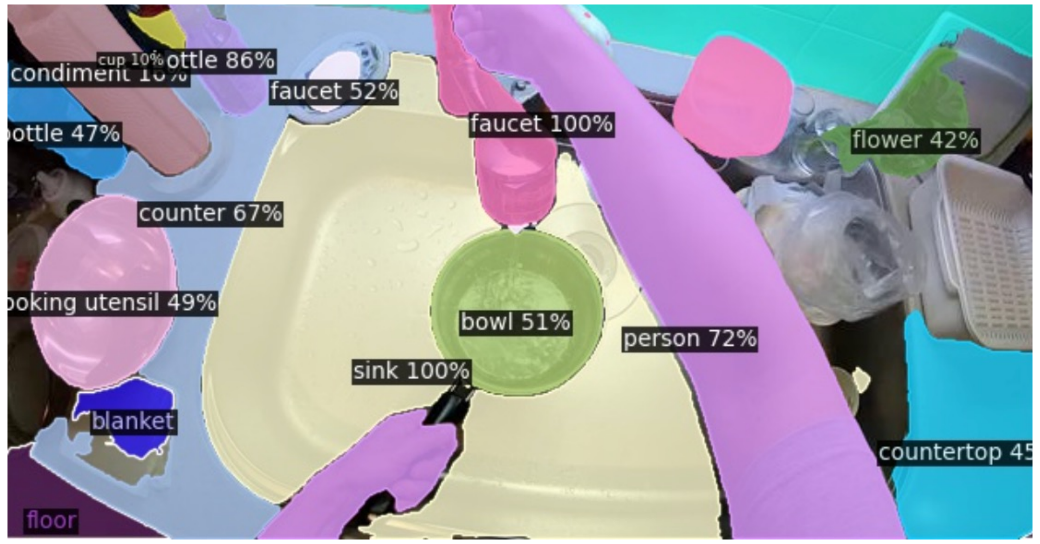
text-only supervision



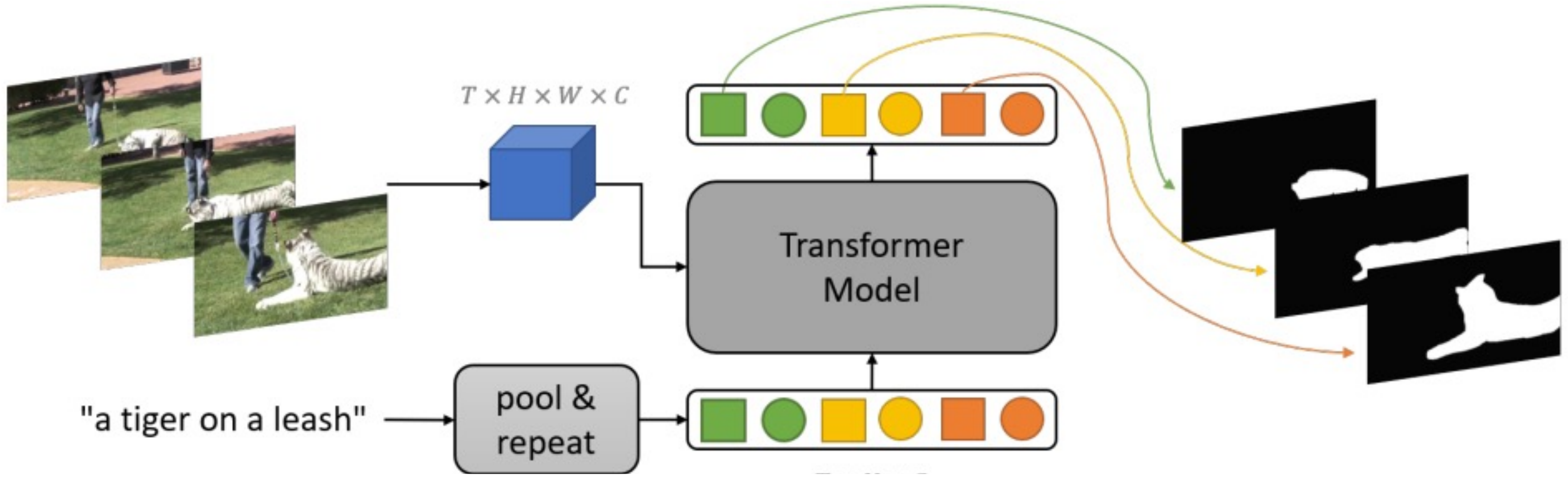
NAFAE [2]

Pixel-level grounding/segmentation

spatial supervision



ODISE [3]



ReferFormer [4]

1. S. Kurita, et al. RefEgo: Referring Expression Comprehension Dataset from First-Person Perception of Ego4D. ICCV 2023.
2. J. Shi, et al. Not All Frames Are Equal: Weakly-Supervised Video Grounding with Contextual Similarity and Visual Clustering Losses. CVPR 2019.
3. J. Xu, et al. ODISE: Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models, CVPR 2023.
4. J. Wu, et al. Language as Queries for Referring Video Object Segmentation. CVPR 2022.

NVOS Challenges in Egocentric Videos

Fine-grained segmentation with object state changes & occlusions

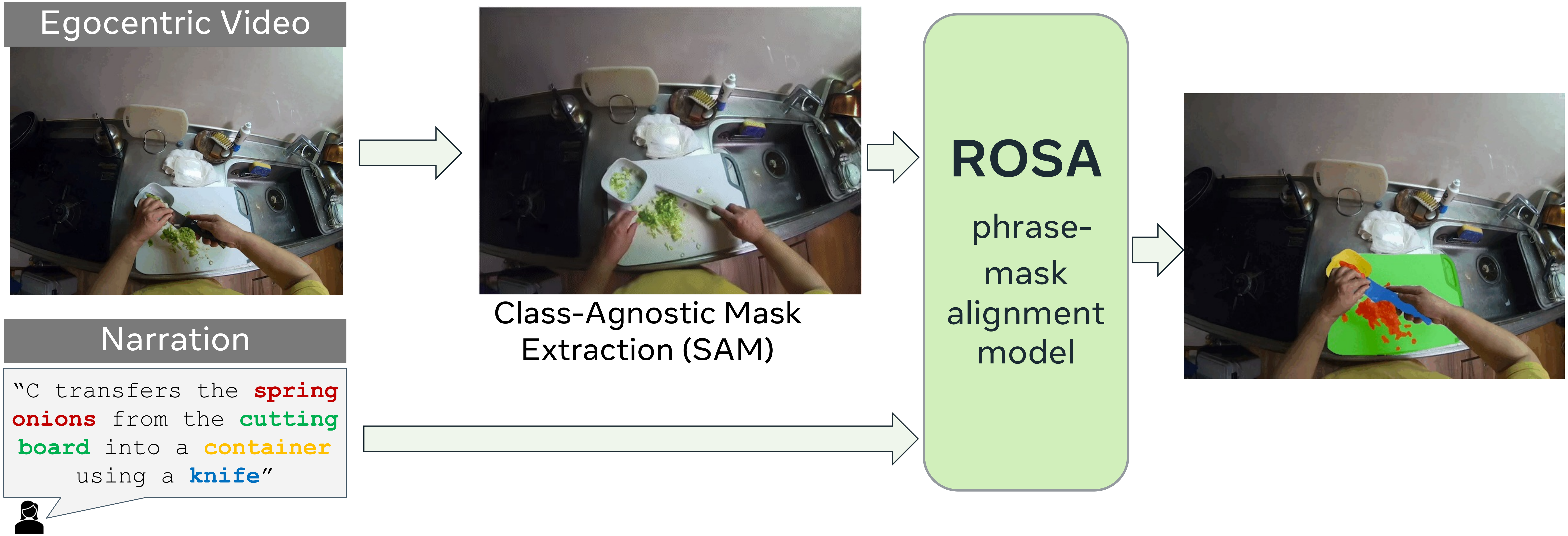
Object instance disambiguation in cluttered scenes

Lack of large-scale datasets with narrations and spatial annotations

- How to train models for NVOS without any spatial annotations? How to evaluate?

Examples from our VISOR-NVOS benchmark

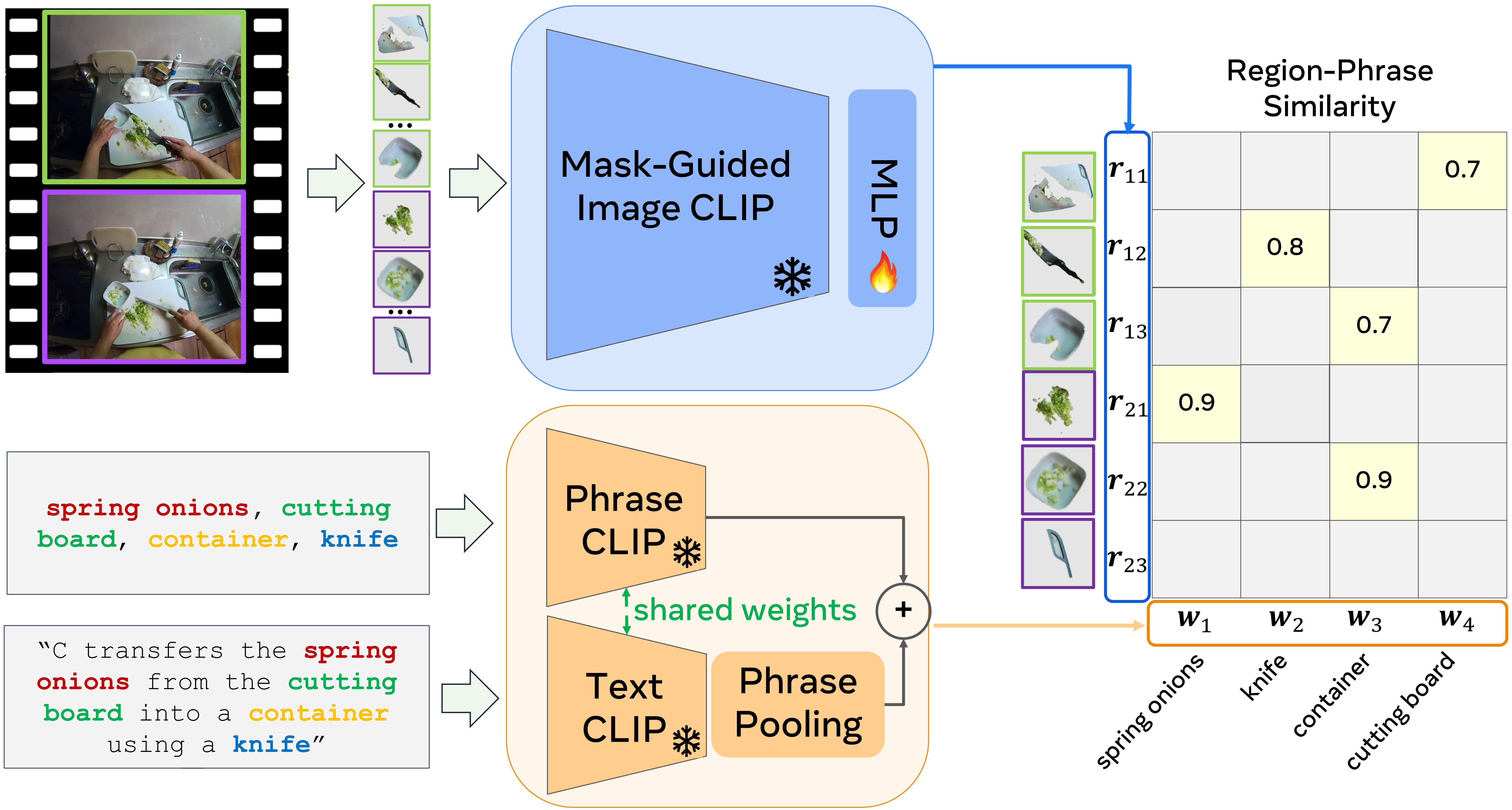
Referred Object-Segment Aligner (ROSA)



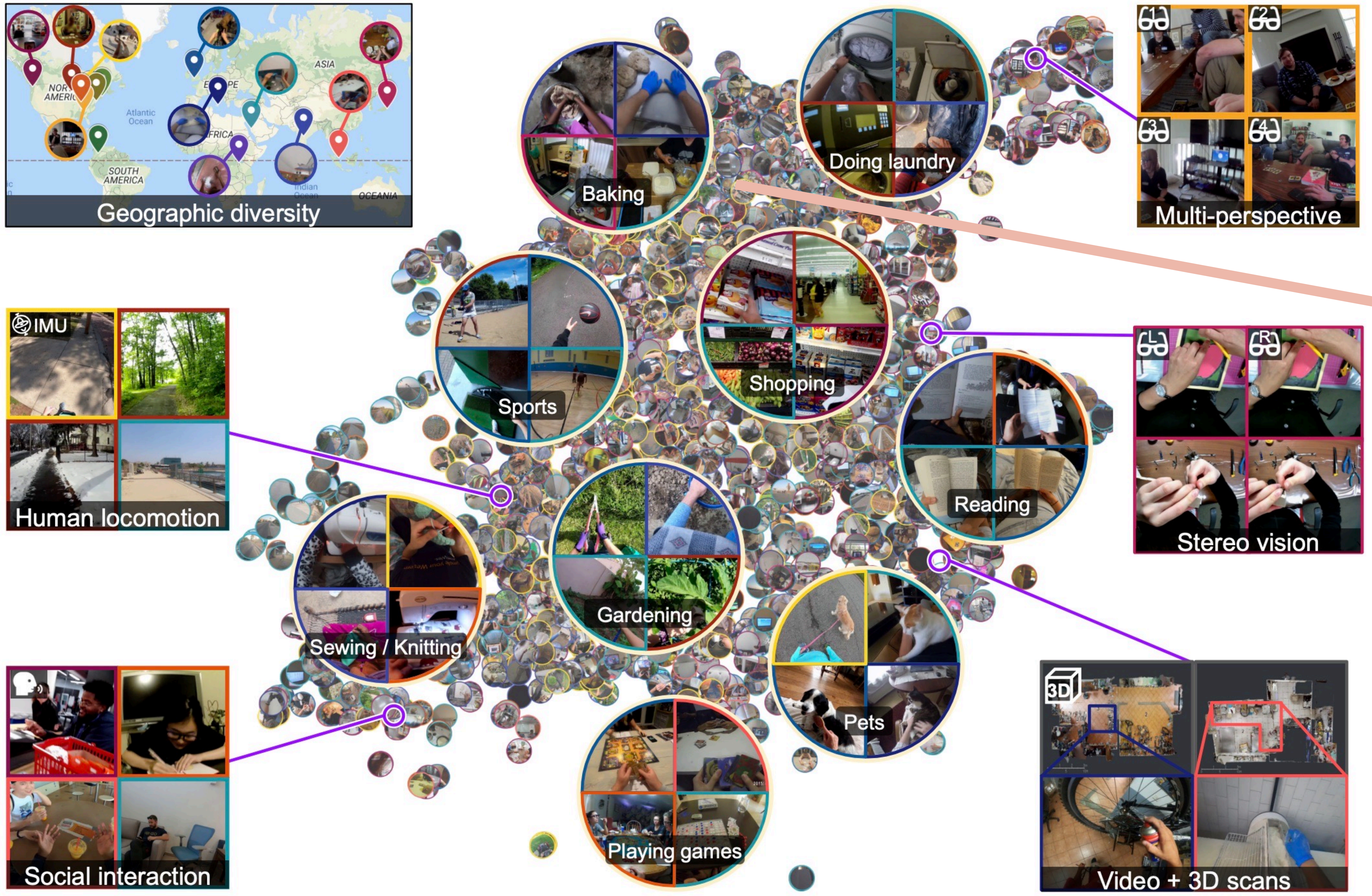
- **NVOS as phrase-mask alignment**: leverage mask proposals from SAM [1]
- **ROSA:** trained with weak supervision of egocentric video narrations
- **VISOR-NVOS** evaluation benchmark: 14k egocentric clips with narrations, 37k referred objects

1. A. Kirillov et al. Segment anything. ICCV 2023.

ROSA Architecture



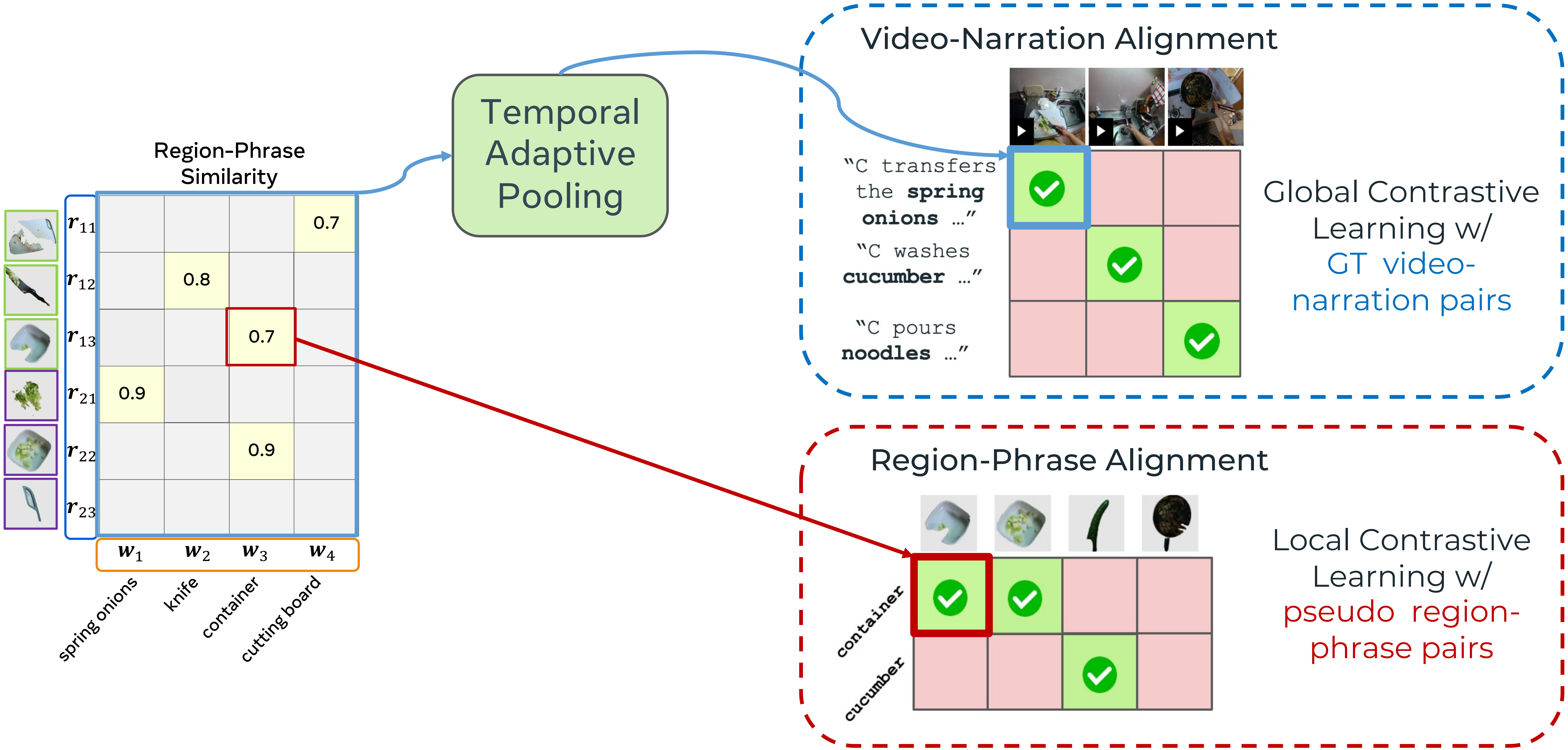
Training: Weak supervision from Ego4D Narrations



Provides dense human narrations for egocentric video

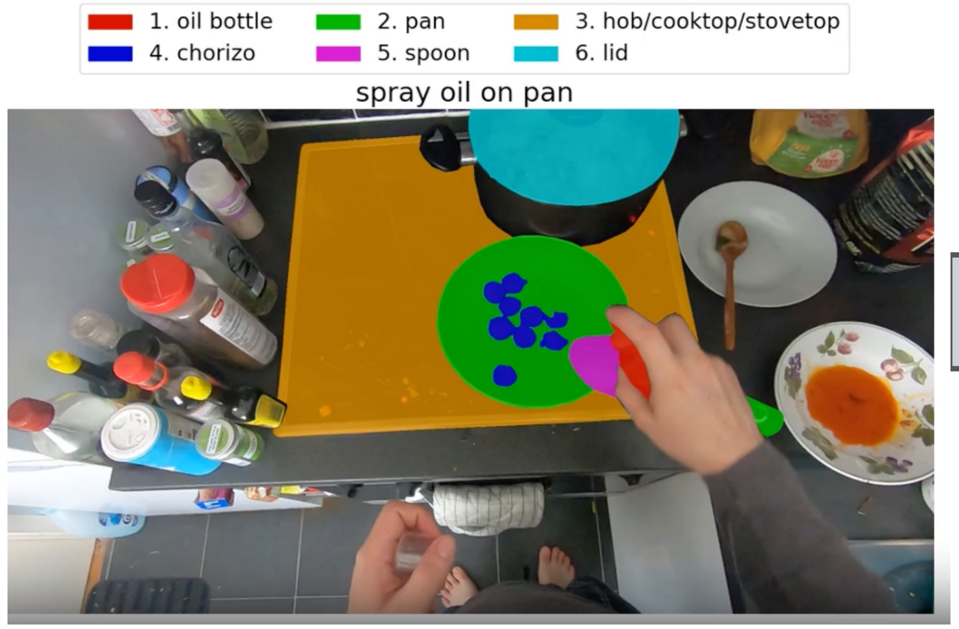
1. K. Grauman et al. Ego4d: Around the world in 3,000 hours of egocentric video. CVPR 2022.
2. S. Ramakrishnan, et al. Naq: Leveraging narrations as queries to supervise episodic memory. CVPR 2023.

Global-Local Contrastive Learning



Evaluation: Introducing VISOR-NVOS

- The first benchmark for narration-based egocentric video object segmentation
- Collected rich narration annotations on top of VISOR segmentation masks (efficient)



Frame with segmentation masks from VISOR

2sec video clip around frame

"C sprays the oil with the [oil bottle]<1> on the [chorizo]<4> in the [pan]<2> placed on the [hob]<3>."

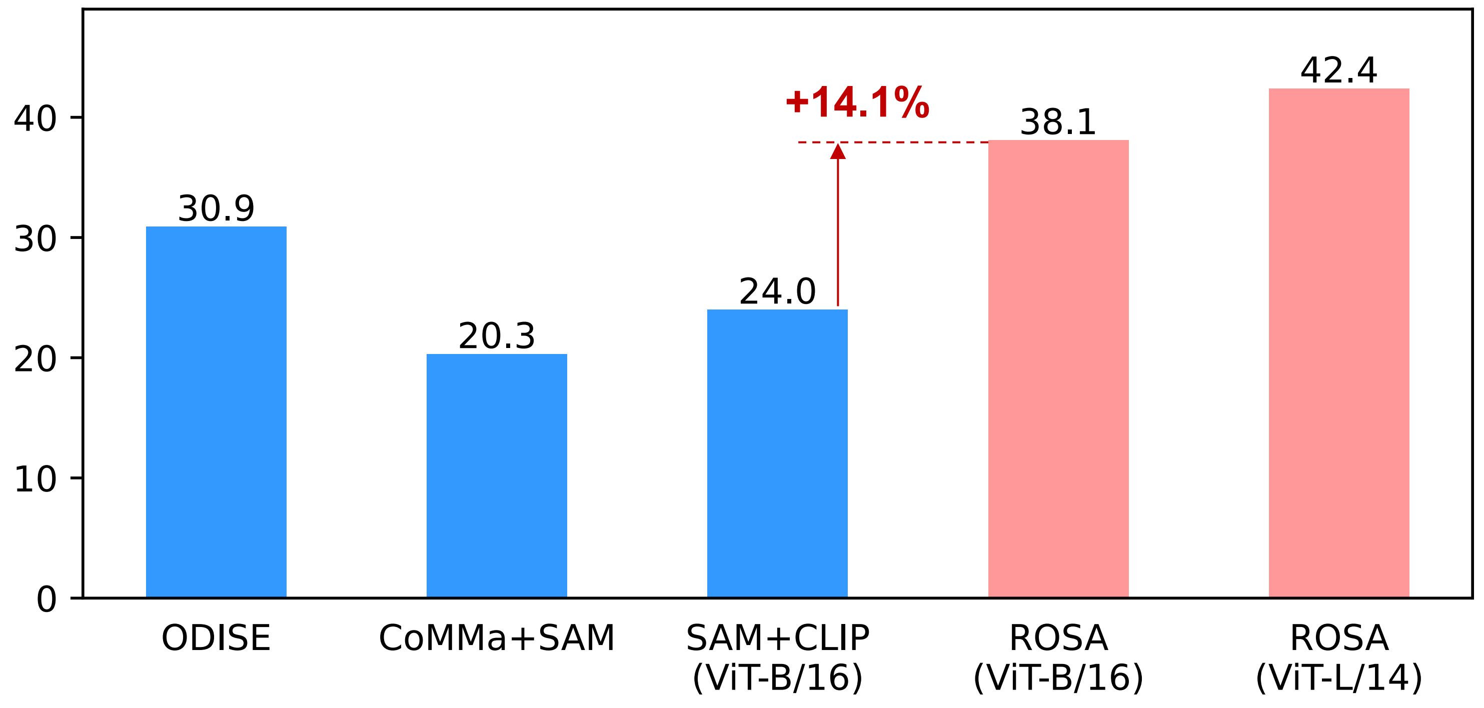
Rich narration with grounded phrases

VISOR-NVOS (val/test): 7.5k/7k clips, 19.6k/17.6k referred objects,
2.54 objects per narration (avg), 12.8 words per narration (avg)

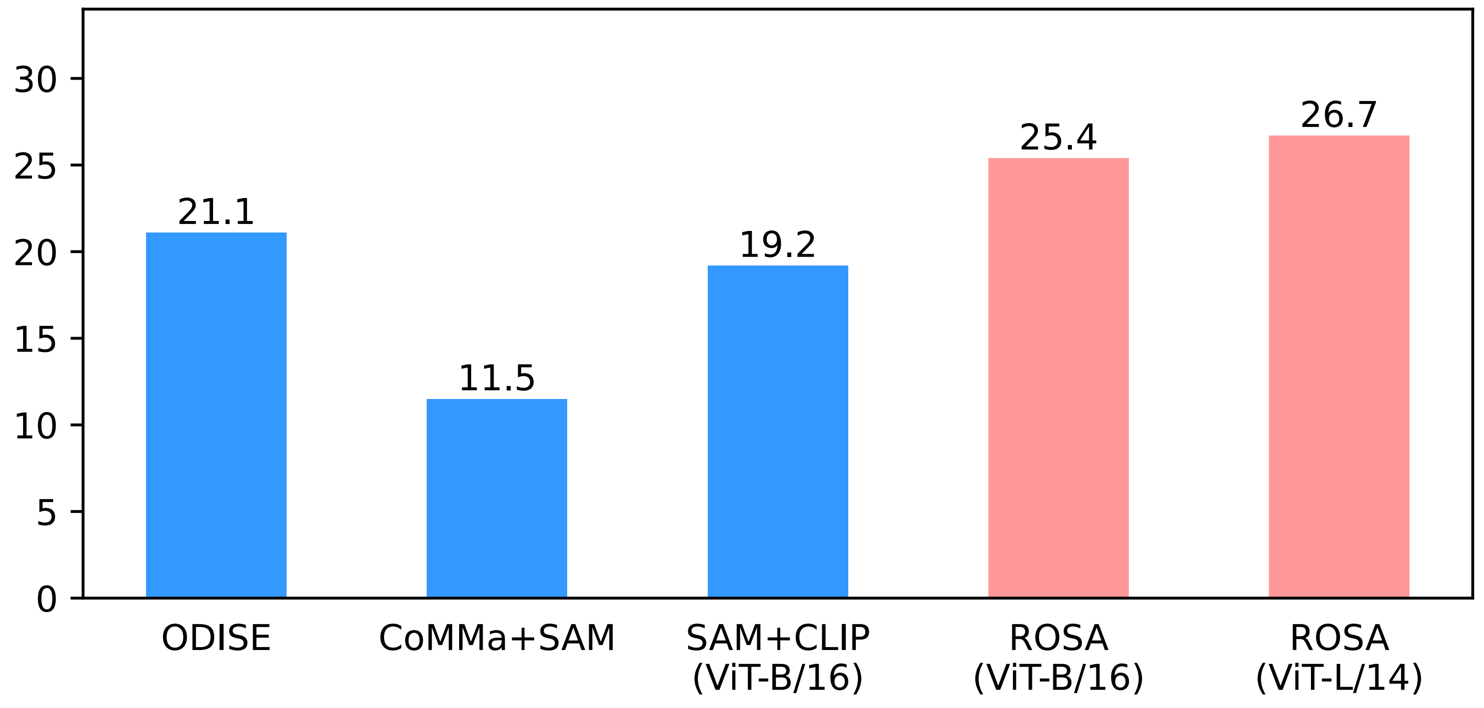
1. A. Darkhalil et al. Epic-kitchens visor benchmark: Video segmentations and object relations. NeurIPS 2022.

Comparison with SOTA

- **Evaluation setup**: compare predicted masks with GT masks at annotated frame(s)
- Our ROSA model outperforms:
  - ODISE [1]: an open-vocabulary object segmentation method trained with labeled segmentation masks
  - CoMMa [2]+SAM: a point-wise grounding method (trained with the same Ego4D narration pairs) followed by point-prompted SAM
  - SAM+CLIP: cropped and masked images + object phrases into CLIP



Performance ($\mathcal{J}$&$\mathcal{F}$) on VISOR-NVOS



Performance ($\mathcal{J}_{ins}$) on VOST [3] (*with action labels as narrations*)

1. J. Xu, et al. Open-vocabulary panoptic segmentation with text-to-image diffusion models. CVPR 2023.
2. R. Tan, et al. Look at what I'm doing: Self-supervised spatial grounding of narrations in instructional videos. NeurIPS 2021.
3. P. Tokmakov, et al. Breaking the "Object" in Video Object Segmentation. CVPR 2023.

Generalization to third-person videos

- Evaluated on YouCook2-BB [1] : third-person videos; bounding boxes
- Perform comparably to supervised training methods on YouCook2-BB

noodles

butter

ginger

lettuce

pan

pan

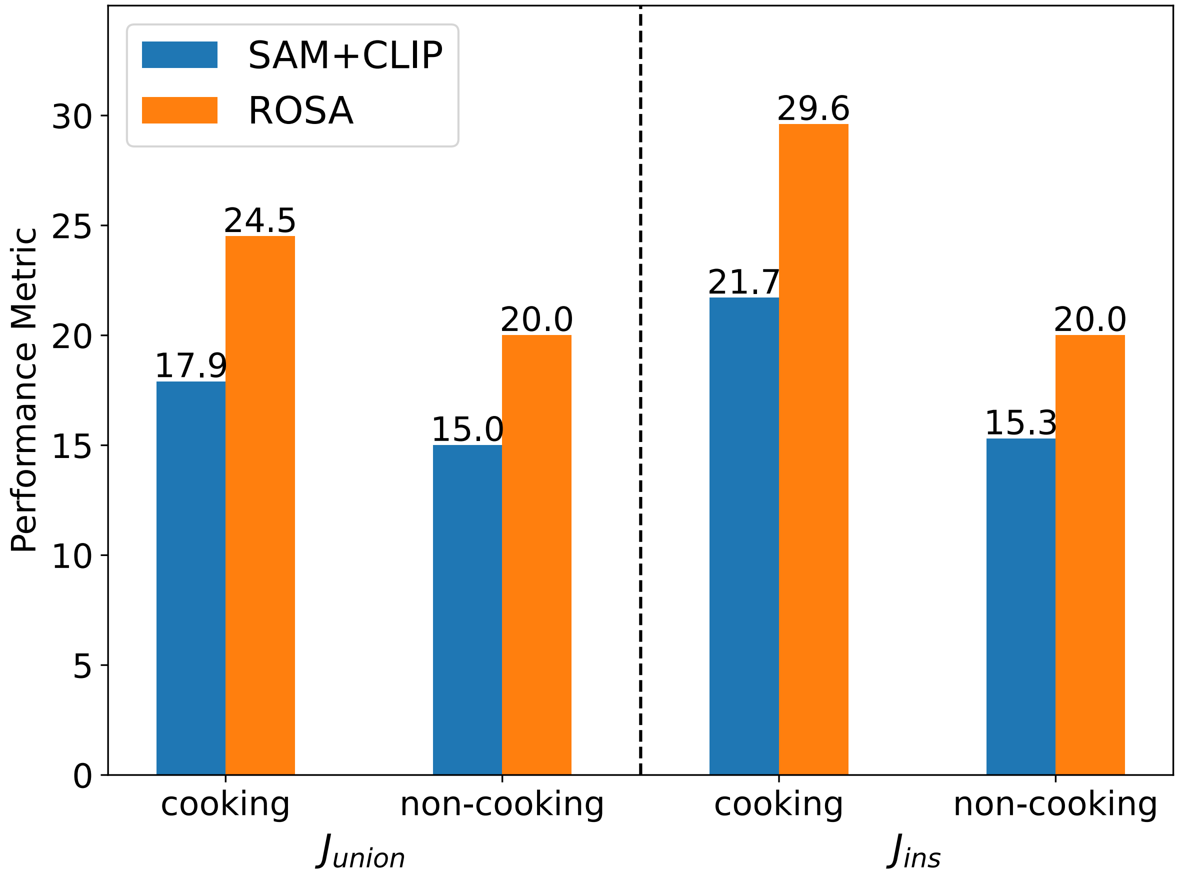
Method	box accuracy	
	macro	micro
Trained on YouCook2		
Zhou et al.	35.08	42.42
NAFAE	40.71	46.33
STVG	41.67	48.22
SCL	**42.80**	**48.60**
Zero-Shot		
CoMMa+SAM	6.63	8.98
Ours	**37.93**	**44.96**

Bounding box evaluation on YouCook2-BB

1. L. Zhou, et al. Weakly- supervised video object grounding from text by loss weighting and object interaction. BMVC 2018

Generalization to non-cooking videos

- VOST [1] contains both cooking and non-cooking videos
- Training on Ego4D cooking videos improves within-domain grounding performance by 6.6% and out-of-domain grounding performance by 5.0%

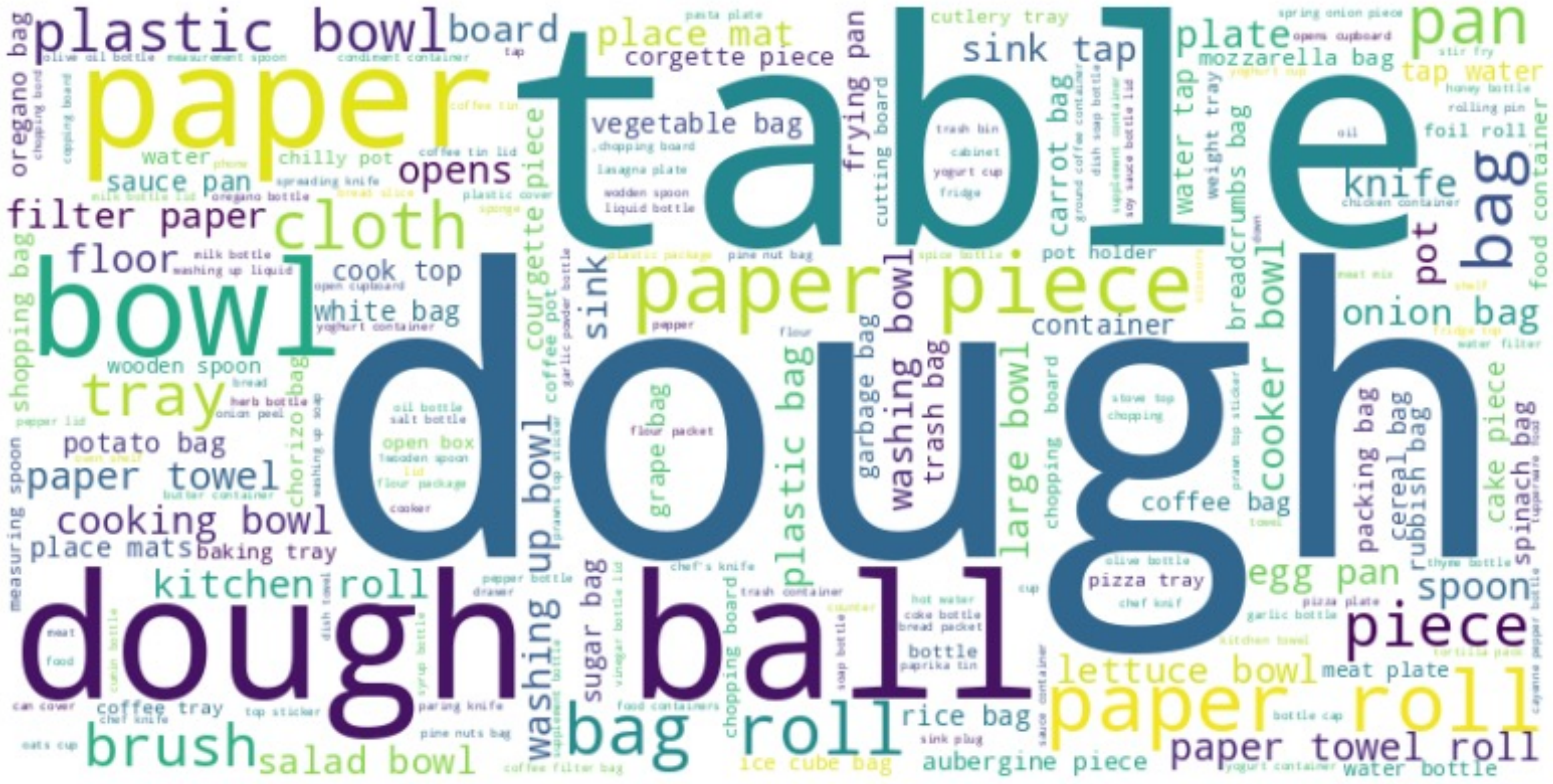


Comparison between ROSA and SAM+CLIP on VOST for cooking and non-cooking videos

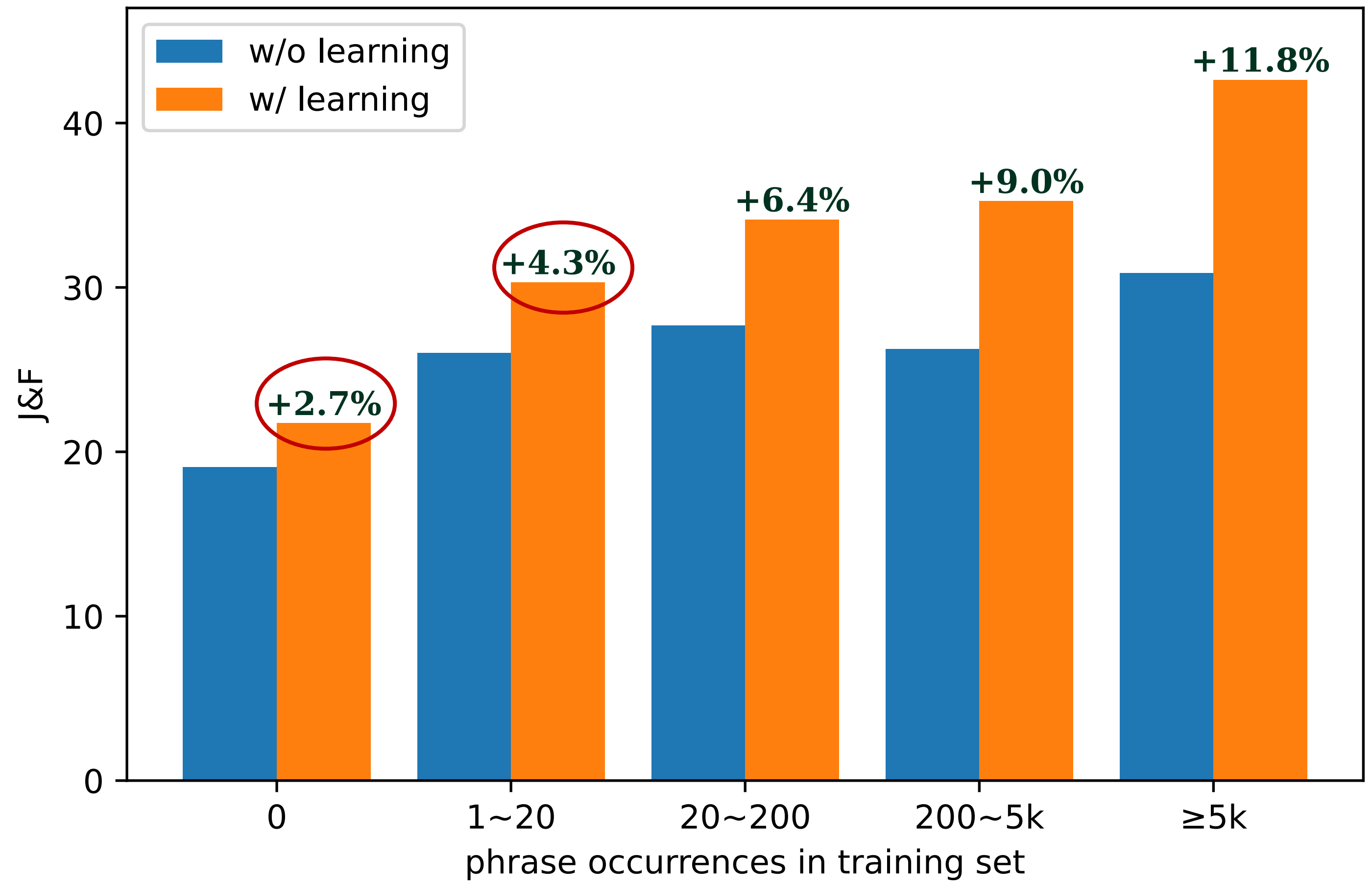
1. P. Tokmakov, et al. Breaking the "Object" in Video Object Segmentation. CVPR 2023.

Generalization to unseen object phrases

- Evaluate performance gain w.r.t. the occurrences of object phrases in training set
- ROSA increases J&F by 2.7% on unseen object phrases and by 4.3% on rare object phrases

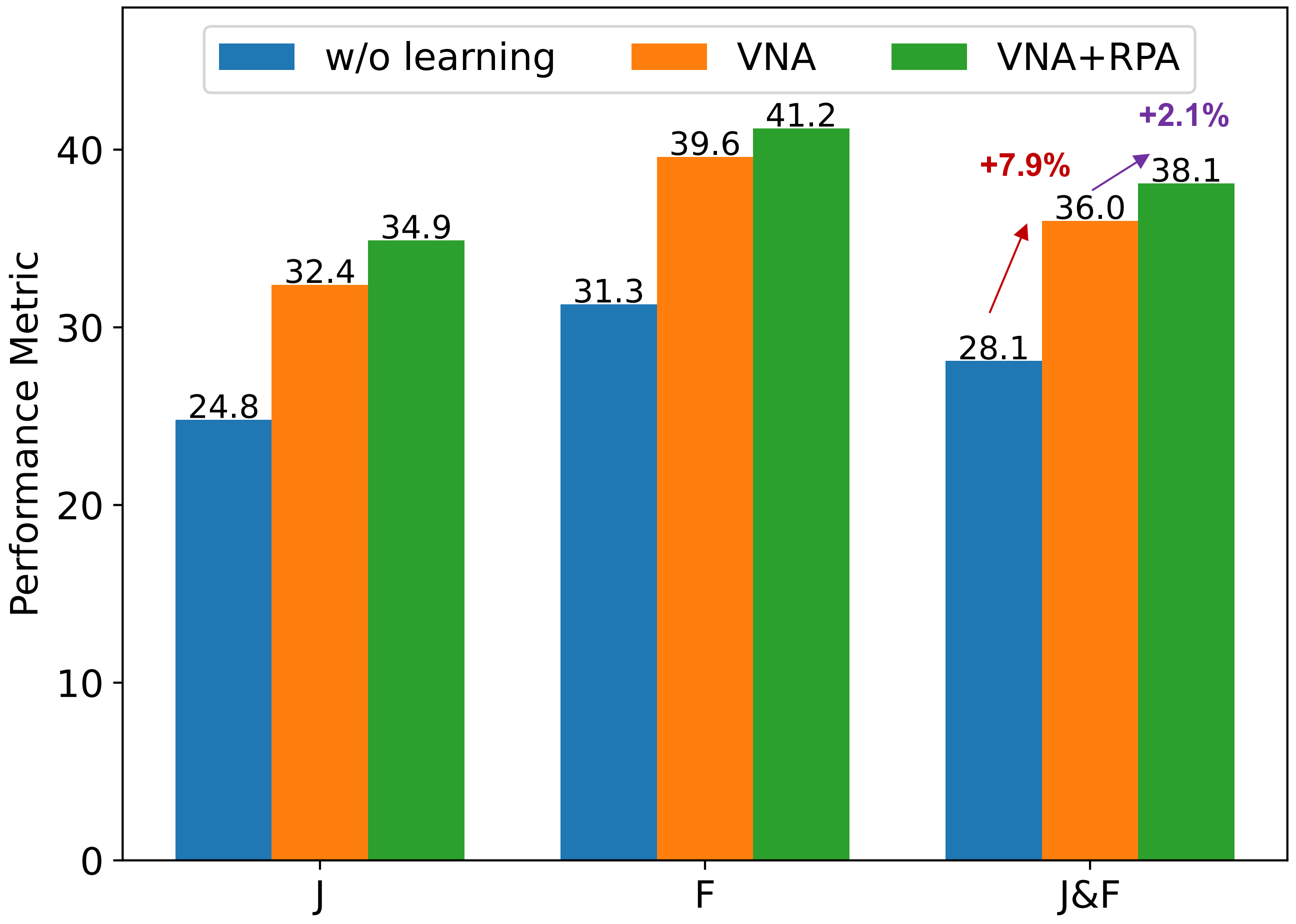


Word cloud of object phrases seen in training set



Performance gain on VISOR-NVOS from ROSA w.r.t. phrase occurrences in the training set

Ablations on VISOR-NVOS: Training Losses



Method	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
w/o learning	25.1	32.3	28.7
fix alignments	29.2	36.4	32.8
update alignments (ours)	**35.0**	**41.9**	**38.5**

Video Narration Alignment (VNA) contrastive loss improves $\mathcal{J}\&\mathcal{F}$ by **7.9%** and Region Phrase Alignment (RPA) contrastive loss further improves by **2.1%**

Dynamically-updated pseudo-labels for region-phrase pairs improve $\mathcal{J}\&\mathcal{F}$ by **5.7%**

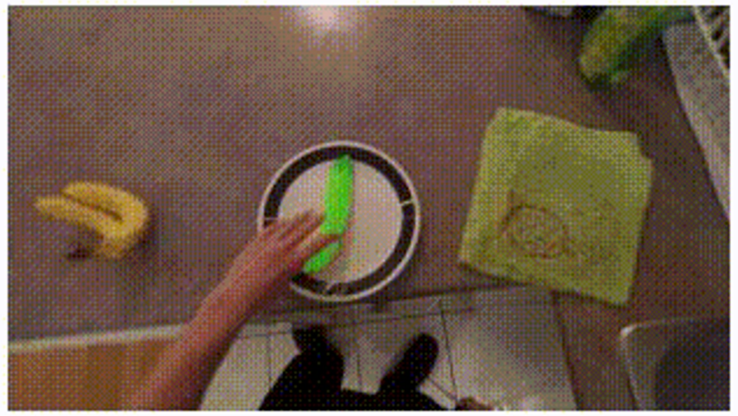
Qualitative Results

"The person mixes **rice** in a **pan** with a **spoon**."

"The person puts **flour** into the **bowl** from the **flour package**."

peel **banana**

SAM+CLIP



ROSA (ours)

plurals

"The person picks **food containers** from the **fridge**."

ambiguity in mask size

"C opens the **fridge** and takes out the **soy milk container** from it."

VISOR-NVOS: Future Directions

Temporal Context

"C opens the drawer and moves the cutlery tray and tries to pick up the spoon."

Active & Inactive Objects

"C puts down the salt bottle and the spoon on the kitchen top and picks up the yeast can in their hands."

Multiple Masks per Object

"C picks up the lids from the hob."

Take home messages

Poster Session

Slot #460 (Arch Exhibit Hall)

- **Task:** pixel-level grounding of referred objects in narrations (NVOS)
- **Method (ROSA)**:
  - Generate mask proposals using SAM and extract object phrases from narrations
  - Obtain **context-aware** representations for mask regions and object phrases using CLIP
  - Learn from text-only supervision via global video-narration alignment and local region-phrase alignment using pseudo-labels
- Introduce **VISOR-NVOS** a benchmark for narration-based egocentric video object segmentation