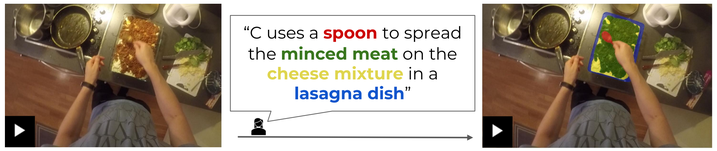
Learning to Segment Referred Objects from Narrated Egocentric Videos

Yuhan Shen[1,2], Huiyu Wang[1], Xitong Yang[1], Matt Feiszli[1], Ehsan Elhamifar[2], Lorenzo Torresani[1], Effrosyni Mavroudi[1]
[1]FAIR, Meta [2]Northeastern University

CVPR JUNE 17-21, 2024 SEATTLE, WA

Overview

- **Narration-based Video Object Segmentation** (NVOS): segment object instances mentioned in narrations for egocentric videos
- **Referred Object-Segment Aligner** (ROSA): a weakly-supervised framework for NVOS without spatial annotations
- **VISOR-NVOS**: an NVOS benchmark with newly-collected video clip narrations and associated segmentation masks

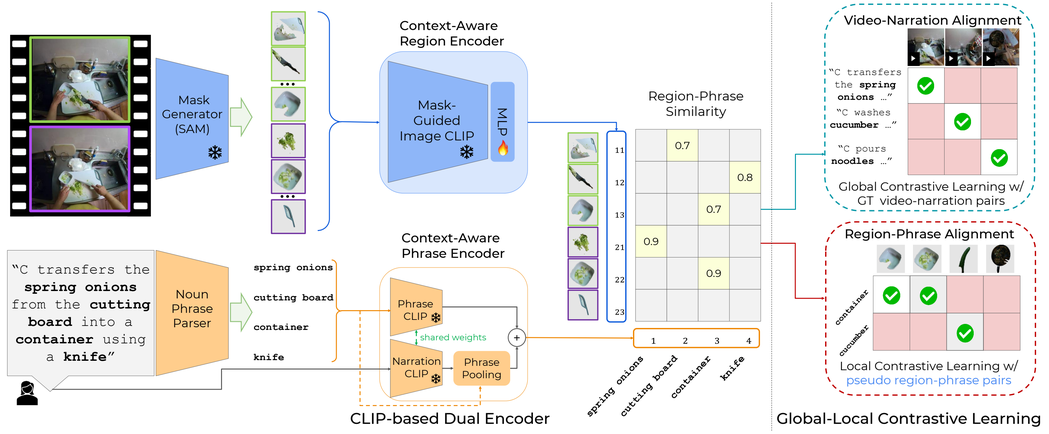


Narration-based Video Object Segmentation (Ours)

Comparison with Related Tasks

References
[1] Kirillov et al. Segment anything. ICCV 2023.
[2] Radford et al. Learning transferable visual models from natural language supervision. ICML 2021.
[3] Darkhalil et al. Epic-kitchens visor benchmark: Video segmentations and object relations. NeurIPS 2022.
[4] Tokmakov et al. Breaking the "Object" in Video Object Segmentation. CVPR 2023.

Referred Object-Segment Aligner (ROSA)



- Generate *mask proposals* for video clips using SAM [1]; extract *object phrases* from narrations
- **CLIP-based Dual Encoder**: obtain *context-aware* representations for segmentation masks and object phrases via pretrained CLIP [2] models
- **Global-Local Contrastive Learning**: contrastive training via global *video-narration alignment* (VNA) and local *region-phrase alignment* (RPA)

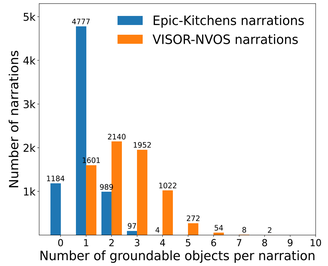
$$\mathcal{L}_{\text{VNA}} = \frac{1}{B}\sum_{i=1}^{B} -\log\frac{e^{\phi(\mathcal{V}_i,\mathcal{S}_i)/\tau}}{\sum_j e^{\phi(\mathcal{V}_i,\mathcal{S}_j)/\tau}} - \log\frac{e^{\phi(\mathcal{V}_i,\mathcal{S}_i)/\tau}}{\sum_j e^{\phi(\mathcal{V}_j,\mathcal{S}_i)/\tau}}$$

$$\mathcal{L}_{\text{RPA}} = -\sum_t \underbrace{\log\frac{e^{g(\tilde{\mathbf{r}}_{tn},\mathbf{w}_n)/\tau}}{\sum_{n'} e^{g(\tilde{\mathbf{r}}_{tn},\mathbf{w}_{n'})/\tau}}}_{\text{region to phrase}} \underbrace{- \log\frac{\sum_t e^{g(\tilde{\mathbf{r}}_{tn},\mathbf{w}_n)/\tau}}{\sum_{n'}\sum_t e^{g(\tilde{\mathbf{r}}_{tn'},\mathbf{w}_n)/\tau}}}_{\text{phrase to region}}$$

VISOR-NVOS Benchmark

- Annotate object-based narrations for video clips from VISOR [3] dataset
- 7,561 validation videos and 7,051 test videos
- 37,170 referred objects with associated segmentation masks
- Average number of groundable objects per narration: 2.54

Narration:
The person uses a [wooden spoon]<5> to stir [pasta]<2> in a [pan]<3> on the [stovetop]<6>.



Evaluation

Quantitative Results

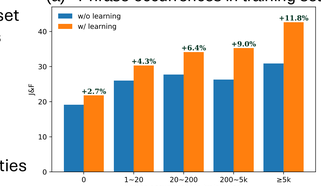
Method	Supervision		VISOR-NVOS			VOST	
	Cross-Modal	Ego4D	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}_{union}$	$\mathcal{J}_{ins}$
SAM upper bound			70.3	75.6	73.0	62.2	65.8
Trained w/ labeled regions							
ODISE	mask-text		29.0	32.8	30.9	17.6	21.1
GroundedSAM	bbox-text		37.3	41.8	39.5	21.8	25.3
Trained w/o labeled regions							
SAM + CLIP (ViT-B/16)	image-text		22.2	25.8	24.0	16.5	19.2
CoMMa + SAM	video-narration	✓	15.3	25.3	20.3	9.4	11.5
ROSA (ViT-B/16)	video-narration	✓	34.9	41.2	38.1	22.2	25.4
ROSA (ViT-L/14)	video-narration	✓	**38.7**	**46.0**	**42.4**	**23.2**	**26.7**

$\mathcal{J}$: pixel-wise IoU; $\mathcal{F}$: F measure of mask boundaries; $\mathcal{J}\&\mathcal{F}$: average of $\mathcal{J}$ and $\mathcal{F}$
VOST [4]: an auxiliary benchmark of grounding objects under complex transforms
$\mathcal{J}_{union}$: $\mathcal{J}$ with the union of all instances; $\mathcal{J}_{ins}$: $\mathcal{J}$ with the best matched instance

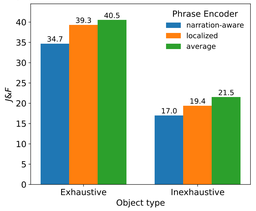
Ablation Studies

(a) More phrase occurrences in training set lead to more gains; ROSA generalizes well to unseen objects (+2.7%)
(b) Adding contexts improves for both exhaustive (+1.2%) and inexhaustive (+2.1%) objects
(c) Both VNA and RPA are effective at learning better region-phrase similarities

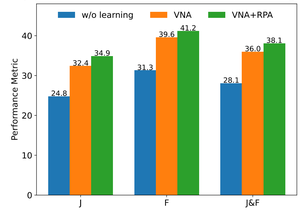
(a) Phrase occurrences in training set



(b) Context in Phrase Encoder



(c) Global-Local Contrastive Learning



Qualitative Results