# 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

## Abstract

*Egocentric videos provide a first-person perspective of the wearer's activities, involving simultaneous interactions with multiple objects. In this work, we propose the task of weakly-supervised Narration-based Video Object Segmentation (NVOS). Given an egocentric video clip and a narration of the wearer's activities, our aim is to segment object instances mentioned in the narration, without using any spatial annotations during training. Existing weakly-supervised video object grounding methods typically yield bounding boxes for referred objects. In contrast, we propose ROSA, a weakly-supervised pixel-level grounding framework learning alignments between referred objects and segmentation mask proposals. Our model harnesses vision-language models pre-trained on image-text pairs to embed region masks and object phrases. During training, we combine (a) a video-narration contrastive loss that implicitly supervises the alignment between regions and phrases, and (b) a region-phrase contrastive loss based on inferred latent alignments. To address the lack of annotated NVOS datasets in egocentric videos, we create a new evaluation benchmark, VISOR-NVOS, leveraging existing annotations of segmentation masks from VISOR alongside 14.6k newly-collected, object-based video clip narrations. Our approach achieves state-of-the-art zero-shot pixel-level grounding performance compared to strong baselines under similar supervision. Additionally, we demonstrate generalization capabilities for zero-shot video object grounding on YouCook2, a third-person instructional video dataset.*

## 1. Introduction

Egocentric videos [6, 12, 42] capture human interactions with the surrounding environment from a first-person perspective. Recent datasets, such as Ego4D [12], have paired such interaction-rich visual signals with manually annotated textual narrations of the camera wearer's activities, catalyzing a surge in research at the intersection of vision and language. These efforts have primarily concentrated on coarse-grained, holistic video-language understanding, such as pairing a video clip with a narration of its content (video-text retrieval) [24, 35, 60], egocentric video question answering [2, 11, 18] and temporal localization of natural language queries [3, 15, 26, 37, 61], often lacking the fine-grained understanding at object level.

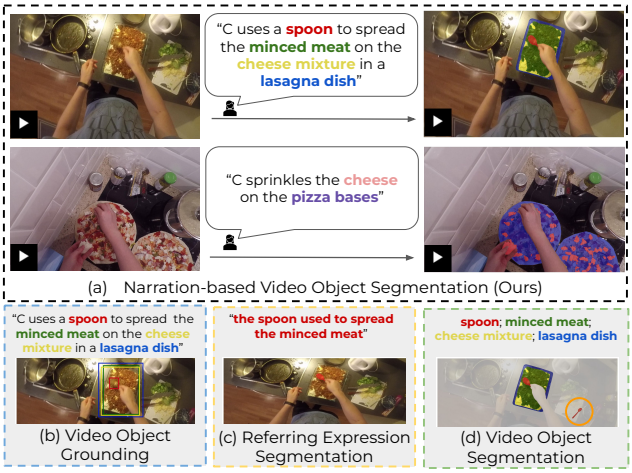


Figure 1. **Overview of the Narration-based Video Object Segmentation (NVOS) task.** Given a short video clip and a narration of the camera wearer's activities, our goal is to predict segmentation masks for each object phrase. This task requires more fine-grained video-language alignment (*pixel-level* and *phrase-level*) compared to popular grounding tasks, such as video object grounding and referring expression segmentation. Frames taken from the Epic-Kitchens [6] dataset.

In this work, we explore a more fine-grained connection between egocentric videos and language by introducing the *Narration-based Video Object Segmentation* (NVOS) task, linking *noun phrases* (objects) in narrations with *segmentation masks* in the video. For example, as shown in Figure 1, given an egocentric video clip and its narration "*C uses a spoon to spread the minced meat on the cheese mixture in a lasagna dish*", our goal is to predict a segmentation mask at each frame for each one of the referred objects (noun phrases): *spoon*, *minced meat*, *cheese mixture*, and *lasagna dish*. Our task requires models to effectively disambiguate

*Work done during an internship at FAIR, Meta.

among multiple instances of the same object class (e.g., predict a segmentation mask for the spoon that the person uses and not for the spoon on the chopping board).

This task is related to, but different from other popular tasks that output segmentation masks, such as *referring expression segmentation* [50, 58] (Figure 1c), which generates a segmentation mask for only one object described in an input query sentence and *open vocabulary segmentation* [23, 38] (Figure 1d), which does not use narration context and is unable to disambiguate object instances of the same class. Besides, most prior approaches that address phrase-level grounding in images [4, 33] and videos [22, 41, 43, 64] were limited to predicting a single point or a bounding box per referred object. However, detecting the bounding boxes (Figure 1b) is not sufficient to localize highly-overlapping, occluded, or non-grid-aligned objects which frequently occur in egocentric procedural videos.

Training models for pixel-level grounding typically requires spatial annotations (e.g., segmentation masks [16, 20] or mouse traces [45]) for the referred objects. However, obtaining precise annotations for every referred object in textual narrations is not only arduous but also prohibitively expensive for large-scale datasets like Ego4D [12]. Hence, in this work we present a weakly-supervised framework for NVOS, trained only with narrated egocentric videos, *without any manual spatial annotations for referred objects*. Our framework, called Referred Object-Segment Aligner (ROSA), leverages recent breakthroughs in class-agnostic segmentation mask generation [21] and casts grounding as a region-phrase alignment problem.

Learning such alignments from uncurated narrated egocentric videos is very challenging, due to cluttered scenes, drastically changing object appearance (e.g., *cheese getting shredded*) and frequent occlusions. To address this, we propose: (a) bootstrapping from powerful joint image-text models, such as CLIP [36], to get a prior about region-phrase alignments, and (b) formulating a global matching score between a video and a narration based on the local matching scores between the masks and noun phrases across different video frames. In particular, we design a *CLIP-based Dual Encoder* for obtaining context-aware embeddings for mask proposals and noun phrases. We also design a *Global-Local Contrastive Learning* framework that operates on ground-truth video-narration pairs and pseudo-labelled region-phrase pairs.

To address the lack of benchmarks for pixel-level grounding in egocentric videos, we introduce a new benchmark, VISOR-NVOS, through manual annotations of narrations referencing segmented objects from the VISOR dataset [7]. This benchmark comprises detailed object-based narrations for 14,612 video clips.

We summarize the contributions of this work as follows:

- We introduce a new task of weakly-supervised NVOS, and establish a new egocentric benchmark, VISOR-NVOS, by enriching the VISOR dataset [7] with detailed object-based narrations.
- We propose ROSA, a weakly-supervised framework that learns to align referred objects in narrations with mask proposals from narrated egocentric videos, without requiring spatial annotations for semantic concepts.
- We achieve state-of-the-art zero-shot NVOS grounding performance in two egocentric video datasets (VISOR-NVOS and VOST [44]) compared to baselines that use similar weak supervision or stronger spatial supervision. Additionally, we showcase generalization ability for zero-shot video object grounding on the YouCook2-BB [64] third-person video dataset.

## 2. Related Work

**Video object grounding** is an active research field in the intersection of vision and language that aims to localize objects referenced in visual descriptions of videos. Prior work has mostly focused on third-person video datasets, such as YouCook2-BB [64] and ActivityNet Entities [65], and on coarse-grained spatial localization of each object with a bounding box [28, 39, 41, 49, 64] or a single point [43]. Proposed models for referring expression or active object grounding [22, 51] in egocentric videos also output rectangular regions. Instead our work explores pixel-level grounding of referred objects in egocentric videos.

Our NVOS task is also closely related to the recently introduced Video Narration Grounding task [45], where the goal is to predict a segmentation mask for each noun in a collection of captions of a third-person video. Our work differs by grounding *noun phrases* on *egocentric* videos and by training models without any spatial annotations for referred objects (while [45] used mouse traces).

**Video object segmentation** (VOS) is the problem of pixel-accurate separation of particular objects from the background in videos. The most related ones to our work are semi-supervised VOS [5, 14, 31, 55]. Given the segmentation masks of the objects in the first one or few frames, semi-supervised VOS aims to segment the target objects through the video. However, those methods merely use the visual modality to track and segment objects while our method segments objects based on the language narrations.

**Open-vocabulary segmentation** aims to segment all object instances of the same category for a list of arbitrary categories described through textual input [9, 17, 23, 52, 57, 59]. Instead, our task aims to precisely identify particular instances of referred objects within the video.

**Image phrase grounding** aims to localize corresponding objects in an image given a phrase in the image's caption [4, 13, 46, 47]. While related, we focus on pixel-wise grounding of phrases in videos than rather than bounding-

box localization in images. To address the unique challenges of video context, we adaptively weigh various frames for establishing video-narration affinity scores during training, which has not been explored in those image-based works. Besides, most of those works apply contrastive learning only at image-caption level, with the exception of [4] which proposes a contrastive learning method at the phrase-bounding box level. Our framework is the first to combine both global and local contrastive learning losses for weakly-supervised region-phrase alignment.

**Vision-language pre-training**, *e.g.*, CLIP [36], has fueled a growing interest in knowledge transfer for various downstream tasks [25, 32, 52, 57]. Prior works [23, 57, 62, 63] have shown that naive ways to leverage pre-trained CLIP to represent bounding boxes or segmentation masks are not feasible as CLIP was pre-trained on image-text pairs without any pixel-level annotation. Our work explores ways of adapting CLIP to represent local regions and phrases, which is an exciting research question [23, 58, 63].

# 3. ROSA: Learning Region-Phrase Alignments

## 3.1. Task Formulation

Given a video-narration pair $(\mathcal{V}, \mathcal{S})$, where the video clip $\mathcal{V}$ consists of $T$ frames $\{I_1, I_2, \cdots, I_T\}$ of size $H \times W$, and the narration $\mathcal{S}$ describes the activities demonstrated in the video and refers to $N$ noun phrases (objects) $(o_1, o_2, \ldots, o_N)$, our goal is to output a set of binary segmentation masks $\mathcal{M}_{tn} \in \{0,1\}^{H \times W}$ for the $n$-th referred object at the $t$-th frame. In this work, we focus on weakly-supervised training, i.e., training only with video-narration pairs without any pixel-level annotation. To achieve this, we extract $M$ mask proposals $\{P_{tm}\}_{t,m=1}^{T,M}$ at each frame using the Segment Anything Model (SAM) [21] which is pre-trained on object segmentation masks *without* object labels. Then we use the video-narration pairs to supervise the learning of a similarity function between each mask proposal and each object phrase.

**Method overview.** Figure 2 gives an overview of our proposed framework, Referred Object-Segment Aligner (ROSA). We propose a *CLIP-based Dual Encoder* to compute context-aware embeddings for mask regions and object phrases (Section 3.2). To train our model without ground-truth region-phrase alignments, we propose a *Global-Local Contrastive Learning* framework (Section 3.3). It optimizes two objectives: (1) a global contrastive objective, contrasting ground-truth positive and negative pairs of video clips and narrations, based on a *Temporal Adaptive Pooling* of region-phrase similarities, and (2) a local contrastive objective, contrasting pseudo-labelled positive and negative pairs of regions and phrases. After learning, grounding is achieved by finding the mask proposal that maximizes the region-phrase similarity (Section 3.4).

## 3.2. CLIP-based Dual Encoder

In order to effectively align phrases with mask proposals, we need discriminative region representations that align well with language and capture contextual nuances. We propose a *Context-aware Region Encoder* to embed each mask proposal, and a *Context-aware Phrase Encoder* to embed each groundable noun phrase (object), by leveraging the CLIP model [36] pretrained on large-scale image-text pairs.

### 3.2.1 Context-Aware Region Encoder

The original Vision Transformer [10] (ViT) from CLIP consists of $L$ Transformer layers and embeds a whole image in its special `CLS` token of each last layer. To extract region-level features per mask proposal of a frame, we propose *Mask-Guided Image CLIP*. We pass the frame through the first $L-1$ layers of ViT, and then compute a `CLS` token per mask proposal by applying a simple attention masking in the last Transformer layer, which ensures that the `CLS` token exclusively attends to the masked region. As this approach shares the computation of all hidden layers except for the last Transformer layer among all mask proposals, it is much more efficient than cropping and masking used in prior works [8, 23, 53]. Besides, as the visual encoder has access to the whole frame, it is able to capture contextual information, making it more suitable for grounding task that needs to disambiguate between object instances of the same class based on context. The final region embedding $\mathbf{r}_{tm} \in \mathbb{R}^D$ for the $m$-th mask proposal at the $t$-th frame is computed by projecting the `CLS` token to a $D$-dimensional vector with a shallow MLP, denoted as $\mathbf{r}_{tm} = RegionEncoder(P_{tm}, I_t)$.

### 3.2.2 Context-Aware Phrase Encoder

We use a language parser (spaCy [1]) to extract the noun phrases in each narration and then apply our Context-Aware Phrase Encoder to obtain a contextualized embedding for each phrase. A straight-forward way for obtaining phrase embeddings using CLIP is to pass the entire narration through its text encoder, and perform average pooling over the tokens corresponding to each referred object (*Phrase Pooling*) to get a *narration-aware phrase embedding* for each object. However, we observe that this embedding is often heavily influenced by the narration context, resulting in similar embeddings for nearby phrases, thus hurting disambiguation that is necessary for our grounding tasks. To address this limitation, we additionally input each object phrase to CLIP's text encoder and use the `EOT` (End of Text) token as a *localized phrase embedding*. The final phrase embedding $\mathbf{w}_n \in \mathbb{R}^D$ for the $n$-th object phrase in the input narration is computed as the average of the *localized phrase embedding* and *narration-aware phrase embedding*,

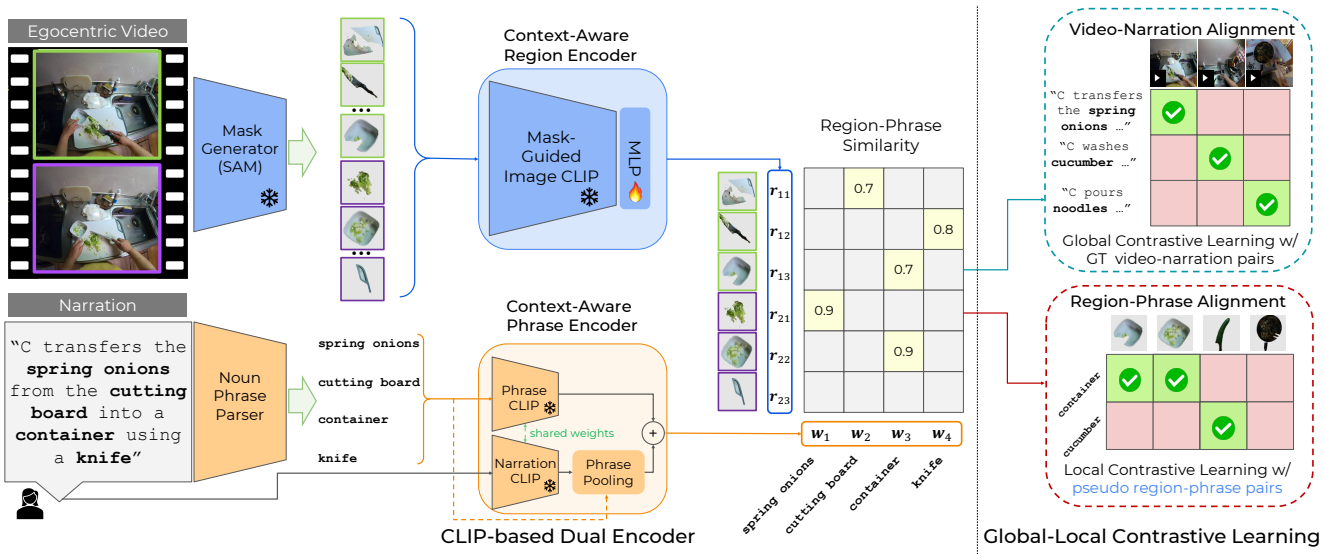


Figure 2. **Framework Overview**. We propose a CLIP-based Dual Encoder to embed mask proposals and noun phrases, and compute pairwise region-phrase similarities. We propose Global-Local Contrastive Learning via ground-truth Video-Narration Alignment and pseudo labelled Region-Phrase Alignment. Frames taken from Ego4D [12].

denoted as $\mathbf{w}_n = PhraseEncoder(o_n, \mathcal{S})$,

### 3.3. Global-Local Contrastive Learning

In the weakly-supervised setting, we need to learn a similarity function between mask regions and object phrases, without ground-truth alignments of region-phrase pairs. We propose a *Global-Local Contrastive Learning* framework to learn the region-phrase similarities using the alignments at both video-narration (global) and region-phrase (local) levels. In Sec. 3.3.1, we detail how to compute the region-phrase similarities and aggregate them into a video-narration affinity score for global contrastive learning. In Sec. 3.3.2, we illustrate how we obtain the pseudo-labelled region-phrase pairs for local contrastive learning.

#### 3.3.1 Video-Narration Alignment

Since our only available supervision is at the video-narration level, we propose a standard contrastive loss for video-narration alignment, which is equipped with a region-based affinity score between videos and narrations. In particular, inspired by Multiple Instance Learning approaches for weakly-supervised phrase grounding [19], we formulate an affinity score $\phi(\mathcal{V}, \mathcal{S})$ between a video and a narration by aggregating individual region-phrase affinity scores. Intuitively, a video-narration pair should have a high affinity score if each referred object in the narration has a matched mask proposal in each frame of the video clip. Next, we detail how we aggregate the region-phrase similarities into frame-phrase affinity scores resulting in an overall affinity score between a video clip and a narration.

**Frame-phrase affinity score.** Given a set of segmentation mask proposals $\{P_{tm}\}_{t,m=1}^{T,M}$ and a set of object phrases $\{o_n\}_{n=1}^{N}$, we first compute the pairwise similarity $g(\mathbf{r}_{tm}, \mathbf{w}_n)$ between each mask proposal embedding $\mathbf{r}_{tm}$ and each noun phrase embedding $\mathbf{w}_n$. In particular, our similarity function $g(\cdot, \cdot)$ is computed as the cosine similarity between the region embedding and phrase embedding, multiplied with the mask confidence score of the corresponding mask proposal. This multiplication prioritizes mask proposals with higher confidence scores when aligning the object phrase with mask proposals. The affinity score between a phrase and a frame is then defined based on the best matching region at each frame $I_t$: $\psi(I_t, o_n) = \max_m g(\mathbf{r}_{tm}, \mathbf{w}_n)$.

**Video-phrase affinity score.** To compute the affinity score at the video-phrase level, we introduce a *Temporal Adaptive Pooling* of the frame-phrase affinity scores $\psi(I_t, o_n)$:

$$\sigma(\mathcal{V}, o_n) = \sum_t \frac{e^{\psi(I_t, o_n)/\alpha}}{\sum_{t'} e^{\psi(I'_t, o_n)/\alpha}} \psi(I_t, o_n). \quad (1)$$

The function involves a weighted sum across all temporal frames, where the weight for each frame is determined by the frame-phrase affinity scores divided by a temperature parameter $\alpha$. We dynamically adjust $\alpha$ during training, following the schedule $\alpha = \alpha_0 \beta^{-s}$, where $\alpha_0$ is a small constant close to 0, $\beta \in (0, 1)$ is a decay index, and $s$ is the iteration step. Intuitively, at the initial stage of training, the model may struggle to effectively segment objects undergoing occlusions or significant transformations. This dynamic adjustment allows the model to prioritize the most confident

mask proposal in a single frame early in training (with low $\alpha$) and gradually shift towards a smoother, more robust aggregation of region representations from multiple frames as training progresses (with a larger $\alpha$).

**Video-narration affinity score.** Finally, we compute the video-narration affinity score by averaging the video-phrase affinity scores over all object phrases: $\phi(\mathcal{V}, \mathcal{S}) = \frac{\sum_n \sigma(\mathcal{V}, o_n)}{N}$, which assumes that all referred objects in the narration should be present in the aligned video clip.

**Video-narration contrastive loss.** Given a training batch of video-narration pairs $\{(\mathcal{V}_i, \mathcal{S}_i)\}_{i=1}^{B}$, where $B$ is the batch size, we use an InfoNCE loss to train the model via weak supervision as follows:

$$\mathcal{L}_{\text{VNA}} = \frac{1}{B}\sum_{i=1}^{B} -\log \frac{e^{\phi(\mathcal{V}_i,\mathcal{S}_i)/\tau}}{\sum_j e^{\phi(\mathcal{V}_i,\mathcal{S}_j)/\tau}} - \log \frac{e^{\phi(\mathcal{V}_i,\mathcal{S}_i)/\tau}}{\sum_j e^{\phi(\mathcal{V}_j,\mathcal{S}_i)/\tau}}, \tag{2}$$

where $\tau$ is a learnable temperature and $\phi(\mathcal{V}_i, \mathcal{S}_j)$ is our custom video-narration affinity score defined above. We freeze the pre-trained CLIP image and text encoders, and optimize the parameters of the MLP to learn region representations better aligned with phrases.

#### 3.3.2 Region-Phrase Alignment

While the aforementioned video-narration contrastive loss is effective at leveraging video-narration supervision, the loss is not directly applied to the alignment between region-phrase pairs. In order to learn a better alignment between mask regions and object phrases, we propose an additional local loss at region-phrase level. To do so, we first get pseudo-labelled alignments for region-phrase pairs. For each phrase embedding $\mathbf{w}_n$, we get the embedding of its best aligned mask proposal at each frame, denoted as $\tilde{\mathbf{r}}_{tn} = \{\mathbf{r}_{tm} | m = \underset{m'}{\text{argmax}}\ g(\mathbf{r}_{tm'}, \mathbf{w}_n)\}$.

**Region-phrase contrastive loss.** Then we apply a region-phrase alignment loss over all object phrases and all aligned mask regions in a batch:

$$\mathcal{L}_{\text{RPA}} = -\sum_t \underbrace{\log \frac{e^{g(\tilde{\mathbf{r}}_{tn},\mathbf{w}_n)/\tau}}{\sum_{n'} e^{g(\tilde{\mathbf{r}}_{tn},\mathbf{w}_{n'})/\tau}}}_{\text{region to phrase}} \underbrace{- \log \frac{\sum_t e^{g(\tilde{\mathbf{r}}_{tn},\mathbf{w}_n)/\tau}}{\sum_{n'}\sum_t e^{g(\tilde{\mathbf{r}}_{tn'},\mathbf{w}_n)/\tau}}}_{\text{phrase to region}}. \tag{3}$$

The first term is a region to phrase (R2P) standard InfoNCE contrastive loss, which treats other object phrases as negative samples for a given mask region. The second term is a phrase to region (P2R) MIL-NCE-type [30] contrastive loss, which considers the aligned regions in all frames as positive pairs for a given object phrase, while treating regions aligned with other object phrases as negative samples. Note that Eq. (3) is the loss for one object phrase, and we sum over all object phrases for training loss. This loss formulation facilitates the learning of robust region-phrase associations and enhances the quality of object segmentation and grounding. In the experiment section, we compare various strategies on selecting the negative samples for P2R contrastive learning and show the advantages of the proposed one. Our final training objective is: $\mathcal{L} = \mathcal{L}_{\text{VNA}} + \lambda \mathcal{L}_{\text{RPA}}$, where $\lambda$ is the weight for region-phrase alignment loss.

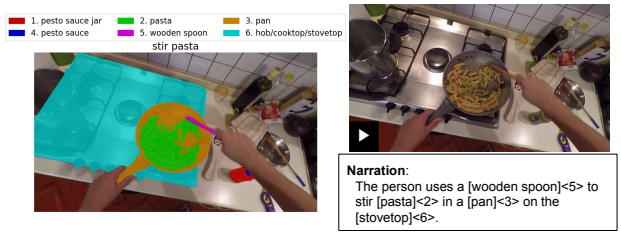


Figure 3. **An example of our annotated narration on the VISOR-NVOS benchmark**. The annotators are instructed to write a detailed narration with [referred object] followed by ⟨object ID⟩.

### 3.4. Inference

We ground the $n$-th noun phrase in the narration by selecting the mask proposal at each frame with the closest embedding in our learned, unified region-phrase embedding space:

$$\mathcal{M}_{tn} = P_{tm^*}, \text{where } m^* = \underset{m}{\text{argmax}}\ g(\mathbf{r}_{tm}, \mathbf{w}_n) \tag{4}$$

## 4. Evaluation Benchmarks for NVOS

While our proposed weakly-supervised framework demonstrates the feasibility of training NVOS models without the need for mask annotation, evaluating the performance of such models is equally critical. Existing egocentric video datasets with segmentation mask annotations for objects, such as VISOR [7], VOST [44] and PACO [38], cannot be directly used for evaluation, since they lack narrations with referred objects associated with the annotated object masks.

**VISOR-NVOS.** To address these limitations, we create a new evaluation benchmark, VISOR-NVOS, by collecting object-based narration annotations for video clips from the VISOR [7] dataset. VISOR provides mask annotations for multiple active objects in egocentric videos of the EPIC-Kitchens [6] dataset. We pair these segmentation masks with narrations, by instructing annotators to describe activities explicitly referring to a set of objects of interest. As shown in Figure 3, provided with a short video clip and a list of objects of interest (with each object name associated with an object ID and a segmentation mask), the annotators are instructed to write a sentence describing the person's

| Method | Supervision | | | VISOR-NVOS | | | VOST | |
|---|---|---|---|---|---|---|---|---|
| | Cross-Modal | Mask Annotation | Ego4D‡ | $\mathcal{J}$ | $\mathcal{F}$ | $\mathcal{J}\&\mathcal{F}$ | $\mathcal{J}_{union}$ | $\mathcal{J}_{ins}$ |
| SAM [21] upper bound | | | | 70.3 | 75.6 | 73.0 | 62.2 | 65.8 |
| *Trained w/ labeled regions* | | | | | | | | |
| ODISE [52] | mask-text | class-specific | | 29.0 | 32.8 | 30.9 | 17.6 | 21.1 |
| GroundedSAM [21, 27] | bbox-text | class-agnostic | | 37.3 | 41.8 | 39.5 | 21.8 | 25.3 |
| *Trained w/o labeled regions* | | | | | | | | |
| SAM [21] + CLIP [36] (ViT-B/16) | image-text | class-agnostic | | 22.2 | 25.8 | 24.0 | 16.5 | 19.2 |
| CoMMa† [43] + SAM [21] | video-narration | class-agnostic | ✓ | 15.3 | 25.3 | 20.3 | 9.4 | 11.5 |
| **ROSA** (ViT-B/16) | video-narration | class-agnostic | ✓ | 34.9 | 41.2 | 38.1 | 22.2 | 25.4 |
| **ROSA** (ViT-L/14) | video-narration | class-agnostic | ✓ | **38.7** | **46.0** | **42.4** | **23.2** | **26.7** |

Table 1. **Comparison of grounding performance between our framework and strong baselines on our NVOS benchmarks**. We report metrics on the test split of our proposed VISOR-NVOS and the validation split of VOST. ‡ whether the model has been fine-tuned on Ego4D. "*Trained with labeled regions*": trained with annotations of bounding boxes or masks; "*Trained without labeled regions*": using video-level or image-level annotations.

activities in the video, linking noun phrases in the sentence with the associated object IDs. In total, we have collected detailed, object-based narrations for 14,612 video clips, including 37,170 referred objects. On average, each narration has 12.79 words, and 2.54 referred objects. The maximum number of referred objects in one narration is 9. We split the annotated videos into a validation set of 7,561 videos and a test set of 7,051 videos, and report the performance on the test set. We do not train our model on any of those videos or narrations. Please see the supplementary for more details about this benchmark.

**VOST.** VOST [44] is an egocentric video object segmentation dataset specifically designed to capture dramatic object transformations. To evaluate models for NVOS on VOST, we use the verb and noun pairs (*e.g.*, "*peel banana*", "*mold clay*") as the narrations, with the nouns serving as the objects to ground. In contrast to VISOR, which encompasses multiple object classes in a single video, VOST contains annotations for a single object class (with one or multiple instances) per video. However, the uniqueness of VOST lies in the complexity of the transformations these objects undergo, *e.g.*, the peeling of bananas. Hence, this benchmark allows us to evaluate the model's performance in grounding objects under complex transformations and its ability to comprehend actions that induce changes in object states.

# 5. Experiments

## 5.1. Experimental Setup

**Training dataset.** We train our model on video-narration pairs sourced from the Ego4D dataset [12]. Ego4D is a massive-scale egocentric video dataset that provides over 3M text narrations of the camera wearer's activities, synchronized with timestamps for $3,670$ hours of video. We use a subset of Ego4D for our experiments, by selecting $250k$ video clips with narrations related to cooking objects. We sample four frames from each one-second clip, with an average of 1.6 narrations available per clip. During each training iteration, we construct video-narration pairs by randomly sampling one narration per video clip.

**Evaluation.** After training on Ego4D, we evaluate on the VISOR-NVOS and VOST benchmarks *without fine-tuning*. Following [7, 34, 54], we use the Jaccard Index ($\mathcal{J}$, also known as Intersection over Union, IoU), the Boundary F-Measure ($\mathcal{F}$), and the average of Jaccard Index and Boundary F-Measure ($\mathcal{J}\&\mathcal{F}$), as the evaluation metrics on VISOR-NVOS. On VOST, following [44], we only use the Jaccard Index as the contours are often not well-defined due to the dramatic transformations and motions involved on this dataset. As we described, VOST does not contain ground-truth alignments between annotated segmentation masks and referred objects. Hence, for some samples with multiple annotated object instances of the same object class, there is uncertainty for grounding evaluation. To mitigate this issue, we report two metrics: $\mathcal{J}_{union}$, i.e., the IoU between the predicted mask and the union of all annotated masks, and $\mathcal{J}_{ins}$, i.e., the IoU between the predicted mask and the most overlapping annotated mask instance.

**Implementation details.** We use SAM [21] to generate mask proposals on each frame using a $32 \times 32$ point grid. We apply Non-Maximum Suppression (NMS) with a threshold of 0.9 to remove redundant mask proposals. Our best model uses a CLIP ViT-L/14 Image Encoder (while for ablations we use ViT-B/16). We set the weight for region-phrase alignment loss $\lambda$ as 0.5 and use three-layer MLP in our Context-Aware Region Encoder. Please refer to the supplementary for more implementation details.

## 5.2. Comparison with the State of the Art

We compare our method with strong baselines as well as state-of-the-art approaches from related tasks (narration grounding [43], open-vocabulary segmentation [52])

| VNA | R2P | P2R | $\mathcal{J}$ | $\mathcal{F}$ | $\mathcal{J}\&\mathcal{F}$ |
|---|---|---|---|---|---|
| | | | 24.8 | 31.3 | 28.1 |
| ✓ | | | 32.4 | 39.6 | 36.0 |
| ✓ | ✓ | | 33.9 | 40.1 | 37.0 |
| ✓ | | ✓ | 33.9 | 40.6 | 37.3 |
| ✓ | ✓ | ✓ | **34.9** | **41.2** | **38.1** |

(a) Impact of various losses.

| Negative samples | $\mathcal{J}$ | $\mathcal{F}$ | $\mathcal{J}\&\mathcal{F}$ |
|---|---|---|---|
| same | 27.5 | 35.0 | 31.3 |
| all | 27.4 | 34.8 | 31.1 |
| other | 26.5 | 34.0 | 30.3 |
| aligned (ours) | **34.9** | **41.2** | **38.1** |

(b) Ablation for negative region sampling.

| | $\mathcal{J}$ | $\mathcal{F}$ | $\mathcal{J}\&\mathcal{F}$ |
|---|---|---|---|
| random frame | 33.3 | 39.8 | 36.6 |
| max pooling | 33.7 | 40.5 | 37.1 |
| softmax | 33.6 | 40.4 | 37.0 |
| adaptive (ours) | **34.9** | **41.2** | **38.1** |

(c) Ablation of temporal aggregation functions.

Table 2. Ablation studies on the Global-Local Contrastive Learning framework on the VISOR-NVOS test split.

adapted for our NVOS task. *Baselines trained without labeled regions*: "SAM+CLIP": For this simple baseline, we use SAM to generate mask proposals, input the cropped and masked images into CLIP image encoder and the object phrases into CLIP text encoder, and select the best matching mask proposal based on the cosine similarity of their embeddings. "CoMMa$^\dagger$+SAM": We adapt CoMMa [43], a state-of-the-art approach for weakly-supervised point-wise grounding of input language queries, for the NVOS task. We train CoMMa$^\dagger$[1] with video-narration pairs (from the same Ego4D subset used to train our model), and during inference use the predicted point for each object phrase to prompt SAM and get the segmentation mask. *Baselines trained with labeled regions:* We also benchmark ODISE [52] and GroundedSAM [21, 27], two state-of-the-art, image-level, open-vocabulary segmentation approaches that are trained with segmentation masks/bounding boxes paired with semantic concepts, thus using stronger supervision than us. Since open-vocabulary segmentation outputs masks for all instances of an object class (*e.g.*, all spoons), we select the mask with highest confidence score and compare it with the ground-truth mask.

As can be seen in Table 1, our model (with CLIP ViT-L/14 backbone) outperforms all compared baselines, including approaches that have been trained with strong spatial supervision (*e.g.*, GroundedSAM is based on Grounding DINO [27], an open-set object detector that has been pre-trained on Objects365 [40], a large dataset with 10M bbox annotations). Interestingly, the simple, out-of-the-box "SAM+CLIP" baseline, which has not been trained on egocentric video data, outperforms "CoMMa$^\dagger$+SAM", corroborating our intuition for using mask proposals and harnessing CLIP pre-training for representing them. Compared to this strong baseline, our model (with the same ViT-B/16 backbone) improves the $\mathcal{J}\&\mathcal{F}$ by **14.1%** on VISOR-NVOS, and improves the $\mathcal{J}_{ins}$ by 6.2% on VOST, which demonstrates the effectiveness of our weakly-supervised framework that learns region-phrase alignments from egocentric video narrations.

We note, however, that the performance of all compared models on our VISOR-NVOS benchmark lags behind the upper bound performance that we could obtain based on the best-matching SAM mask proposal (reported in the first row of Table 1), highlighting the challenging nature of our proposed task and benchmark and the need for further research on learning better region-phrase alignments, *e.g.*, by better modeling temporal consistency.

[1]We use CoMMa$^\dagger$ to denote the model we trained on Ego4D to differentiate with the original CoMMa model trained on HowTo100M.

### 5.3. Ablation Studies

**What is the effect of each component of our proposed training objective?** We start by reporting the performance of our CLIP-based Dual Encoder before weakly-supervised training (*i.e.*, we replace the MLP of our region encoder with the identity mapping and just use CLIP weights for our encoders) in the first row of Table 2a. Adding the MLP and fine-tuning it on Ego4D with our Global Contrastive Objective, that encourages video-narration alignment (VNA), leads to a significant performance improvement in $\mathcal{J}\&\mathcal{F}$ (from 28.1% to 36.0%, rows 1 and 2). Adding our novel region-phrase alignment loss terms, *i.e.*, region to phrase (R2P) and phrase to region (P2R) losses (Eq. (3)), further boosts performance. Both loss terms contribute to the improvement, and their combination improves $\mathcal{J}\&\mathcal{F}$ by 2.1%, suggesting that our proposed loss function is able to learn a better region-phrase alignment by directly applying supervision on region-phrase pairs, instead of just contrasting video-narration pairs, as commonly done in prior work [64].

**How to select negative region samples for region-phrase contrastive learning?** As can be seen in Table 2b, using regions that are *aligned* with the other phrases in the batch (based on the pseudo-labeled alignments) is the only strategy that does not degrade the performance. We compare it with using the rest of regions from the *same* video, using the rest of regions from *all* the videos in the batch, and using all regions from *other* videos (as proposed for weakly-supervised image phrase grounding [4]). We conjecture that our proposed strategy avoids the issues of "fake negative" samples when the negative mask regions are also relevant to the object phrase if multiple mask proposals are overlapped, and "easy negative" samples when the mask regions are background regions irrelevant to active objects.

**What is the effect of the temporal aggregation module?** As we discussed, we aggregate region-phrase similarities from multiple frames to obtain a global video-narration

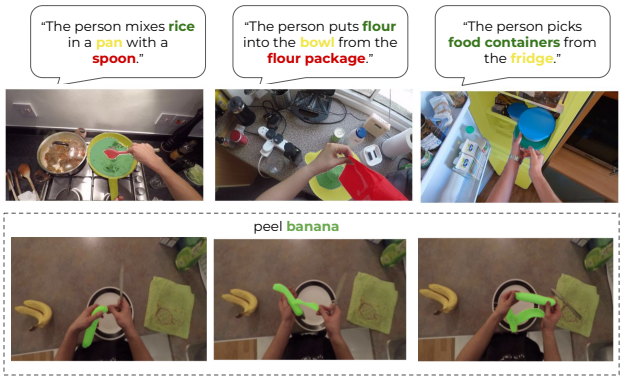


Figure 4. **Qualitative Results**. Top: three examples from VISOR-NVOS. Bottom: an example from VOST showing the model's robustness to objects undergoing dramatic transformations.

affinity score for applying our global contrastive objective. Our proposed aggregation module starts by focusing on one frame for each phrase (the one with the best matching region) and then adaptively starts encouraging alignment with the regions of other frames. As can been seen in Table 2c, this approach (*adaptive*) better handles occlusions and motion blur and facilitates training compared to: (a) randomly selecting a frame of the video clip during each iteration (*random frame*), (b) using the frame with the highest frame-phrase affinity score (*max*), or (c) not adapting the frame weights (*softmax*).

### 5.4. Qualitative Results

In Figure 4, we show some qualitative results. On the top, we show three examples from VISOR-NVOS. In the first example, our model is able to segment the referred objects in a complex environment featuring multiple spoons and pans. In the second example, the model successfully distinguishes between "flour" and "flour package". The third example highlights some limitations of our model, including ambiguity in mask size where the model only segments a portion of the fridge instead of the entire unit, and challenges in understanding plurals (the narration mentions "containers" but it only grounds one container). The bottom section showcases the model's predictions on three frames from a VOST video depicting the peeling of a banana. Remarkably, the model accurately segments the banana while the appearance of banana has changed significantly.

### 5.5. Exocentric Video Object Grounding

In Table 3, to further assess the generality of our model, we conduct evaluations on the third-person video object grounding dataset, YouCook2-BB [64], which includes bounding box annotations for referred objects. We convert the segmentation masks output by our model into bounding boxes for evaluation. Recall that our model is trained on egocentric videos with a goal to align object phrases with segmentation masks. Despite a huge domain shift, our model achieves comparable performance with prior works [41, 48, 56, 64] that are trained on YouCook2 using bounding box proposals (Table 3a). In addition, we adopt the evaluation setup from CoMMa [43] to evaluate our model using point accuracy (see supplementary for more details). We use the center of the bounding box from our segmentation mask as the detected point. While CoMMa is trained on exocentric videos with a similar scale (HowTo100M [29]) for point-based grounding, our model outperforms CoMMa by a significant margin of 10.10% (Table 3b).

| Method | box accuracy | |
|---|---|---|
| | macro | micro |
| Trained on YouCook2 | | |
| Zhou et al. [64] | 35.08 | 42.42 |
| NAFAE [41] | 40.71 | 46.33 |
| STVG [56] | 41.67 | 48.22 |
| SCL [48] | **42.80** | **48.60** |
| Zero-Shot | | |
| CoMMa†+SAM | 6.63 | 8.98 |
| Ours | **37.93** | **44.96** |

(a) bounding box evaluation

| Method | point acc. |
|---|---|
| CoMMa† | 53.56 |
| CoMMa [43] | 59.25 |
| Ours | **69.35** |

(b) point prediction evaluation

Table 3. Object grounding evaluation on YouCook2-BB.

## 6. Limitations

A limitation of our framework is that we perform inference on each frame separately, neglecting temporal information. Integrating temporal context into our model has the potential to enhance performance and alleviate ambiguities, which is a promising direction for future work. Furthermore, we focused our training and evaluation on the cooking domain. In the supplementary, we show that our weakly-supervised training also improves grounding on non-cooking videos, such as DIY, but a performance gap still exists between the two domains.

## 7. Conclusion

We have explored the task of Narration-based Video Object Segmentation and established a new evaluation benchmark, VISOR-NVOS. Our proposed framework, ROSA, has effectively learnt the alignment between referred objects and segmentation masks using weak supervision of video-narration alignments in egocentric videos. Our approach achieved state-of-the-art zero-shot pixel-level grounding performance on two egocentric video datasets compared to strong baselines.

**Acknowledgements.** We thank Tushar Nagarajan, Triantafyllos Afouras, Yale Song, Austin Miller, and Jiabo Hu for helpful discussions and invaluable engineering support.

## References


[1] spaCy. https://spacy.io/. 3

[2] Leonard Bärmann and Alex Waibel. Where did i leave my keys? - episodic-memory-based question answering on egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1560–1568, 2022. 1

[3] Wayner Barrios, Mattia Soldan, Alberto Mario Ceballos-Arroyo, Fabian Caba Heilbron, and Bernard Ghanem. Localizing moments in long video via multimodal guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13667–13678, 2023. 1

[4] Keqin Chen, Richong Zhang, Samuel Mensah, and Yongyi Mao. Contrastive learning with expectation-maximization for weakly supervised phrase grounding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8549–8559, 2022. 2, 3, 7

[5] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022. 2

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4125–4141, 2020. 1, 5

[7] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. *Advances in Neural Information Processing Systems*, 35:13745–13758, 2022. 2, 5, 6

[8] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, 2022. 3

[9] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary panoptic segmentation with maskclip. *arXiv preprint arXiv:2208.08984*, 2022. 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 3

[11] Chenyou Fan. Egovqa-an egocentric video question answering benchmark dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1

[12] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1, 2, 4, 6

[13] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision*, pages 752–768. Springer, 2020. 2

[14] Lingyi Hong, Wenchao Chen, Zhongying Liu, Wei Zhang, Pinxue Guo, Zhaoyu Chen, and Wenqiang Zhang. Lvos: A benchmark for long-term video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13480–13492, 2023. 2

[15] Zhijian Hou, Lei Ji, Difei Gao, Wanjun Zhong, Kun Yan, Chao Li, Wing-Kwong Chan, Chong-Wah Ngo, Nan Duan, and Mike Zheng Shou. Groundnlq@ ego4d natural language queries challenge 2023. *arXiv preprint arXiv:2306.15255*, 2023. 1

[16] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 108–124. Springer, 2016. 2

[17] Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7031, 2022. 2

[18] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. *Advances in Neural Information Processing Systems*, 35: 3343–3360, 2022. 1

[19] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *Advances in neural information processing systems*, 27, 2014. 4

[20] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14*, pages 123–141. Springer, 2019. 2

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2, 3, 6, 7

[22] Shuhei Kurita, Naoki Katsura, and Eri Onami. Refego: Referring expression comprehension dataset from first-person perception of ego4d. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15214–15224, 2023. 2

[23] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 2, 3

[24] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z XU, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language

pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022. 1

[25] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15305–15314, 2023. 3

[26] Bei Liu, Sipeng Zheng, Jianlong Fu, and Wen-Huang Cheng. Anchor-based detection for natural language localization in ego-centric videos. In *2023 IEEE International Conference on Consumer Electronics (ICCE)*, pages 01–04. IEEE, 2023. 1

[27] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 6, 7

[28] Effrosyni Mavroudi and René Vidal. Weakly-supervised generation and grounding of visual descriptions with conditional generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15544–15554, 2022. 2

[29] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019. 8

[30] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 5

[31] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9226–9235, 2019. 2

[32] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 3

[33] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 2

[34] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation, 2018. 6

[35] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5285–5297, 2023. 1

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 6

[37] Santhosh Kumar Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. Naq: Leveraging narrations as queries to supervise episodic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6694–6703, 2023. 1

[38] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7141–7151, 2023. 2, 5

[39] Arka Sadhu, Kan Chen, and Ram Nevatia. Video object grounding using semantic roles in language description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10417–10427, 2020. 2

[40] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 7

[41] Jing Shi, Jia Xu, Boqing Gong, and Chenliang Xu. Not all frames are equal: Weakly-supervised video grounding with contextual similarity and visual clustering losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10444–10452, 2019. 2, 8

[42] Yale Song, Gene Byrne, Tushar Nagarajan, Huiyu Wang, Miguel Martin, and Lorenzo Torresani. Ego4d goal-step: Toward hierarchical understanding of procedural activities. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 1

[43] Reuben Tan, Bryan Plummer, Kate Saenko, Hailin Jin, and Bryan Russell. Look at what i'm doing: Self-supervised spatial grounding of narrations in instructional videos. *Advances in Neural Information Processing Systems*, 34:14476–14487, 2021. 2, 6, 7, 8

[44] Pavel Tokmakov, Jie Li, and Adrien Gaidon. Breaking the" object" in video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22836–22845, 2023. 2, 5, 6

[45] Paul Voigtlaender, Soravit Changpinyo, Jordi Pont-Tuset, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with video localized narratives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2461–2471, 2023. 2

[46] Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. Improving weakly supervised visual grounding by contrastive knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14090–14100, 2021. 2

[47] Qinxin Wang, Hao Tan, Sheng Shen, Michael Mahoney, and Zhewei Yao. Maf: Multimodal alignment framework for weakly-supervised phrase grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2030–2038, 2020. 2

[48] Wei Wang, Junyu Gao, and Changsheng Xu. Weakly-supervised video object grounding via stable context learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 760–768, 2021. 8

[49] Wei Wang, Junyu Gao, and Changsheng Xu. Weakly-supervised video object grounding via causal intervention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3933–3948, 2022. 2

[50] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. *arXiv preprint arXiv:2201.00487*, 2022. 2

[51] Te-Lin Wu, Yu Zhou, and Nanyun Peng. Localizing active objects from egocentric vision with symbolic world knowledge. *arXiv preprint arXiv:2310.15066*, 2023. 2

[52] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 2, 3, 6, 7

[53] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022. 3

[54] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark, 2018. 6

[55] Linjie Yang, Yanran Wang, Xuehan Xiong, Jianchao Yang, and Aggelos K Katsaggelos. Efficient video object segmentation via network modulation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6499–6507, 2018. 2

[56] Xun Yang, Xueliang Liu, Meng Jian, Xinjian Gao, and Meng Wang. Weakly-supervised video object grounding by exploring spatio-temporal contexts. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1939–1947, 2020. 8

[57] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2, 3

[58] Seonghoon Yu, Paul Hongsuck Seo, and Jeany Son. Zero-shot referring image segmentation with global-local context features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19456–19465, 2023. 2, 3

[59] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1020–1031, 2023. 2

[60] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597, 2023. 1

[61] Sipeng Zheng, Qi Zhang, Bei Liu, Qin Jin, and Jianlong Fu. Exploring anchor-based detection for ego4d natural language query. *arXiv preprint arXiv:2208.05375*, 2022. 1

[62] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 3

[63] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 3

[64] Luowei Zhou, Nathan Louis, and Jason J Corso. Weakly-supervised video object grounding from text by loss weighting and object interaction. *BMVC*, 2018. 2, 7, 8

[65] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. Grounded video description. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6578–6587, 2019. 2