Learning to Segment Referred Objects from Narrated Egocentric Videos

Supplementary Material

1. Implementation Details

Mask proposal generator. We generate mask proposals on each frame by applying SAM (ViT-H) [6] on a 32×32 point grid, resulting in 3,072 mask proposals. We apply Non-Maximum Suppression (NMS) with a mask IoU threshold of 0.9 to remove redundant proposals. If more than $1k$ mask proposals remain after NMS filtering, we keep the top $1k$ proposals based on their confidence scores predicted by SAM. During training, we randomly sample 200 mask proposals for each image. We use a 2D Gaussian distribution to model the center positions of all mask proposals, and sample the proposals based on the probability distribution. We empirically observe that this strategy performs slightly better than uniform sampling, as this distribution gives preference to the mask proposals near the center, which are more likely to represent relevant objects in egocentric videos.

Attention masking in Region Encoder. The input resolution of all models is 224×224. For each mask proposal, we first rescale it to the same resolution of the feature maps, *e.g.*, 14×14 for ViT-B/16 model, and then apply attention masking to ensure that the `CLS` token exclusively attends to the masked region.

Noun phrase parser. To get the object phrases on Ego4D, we use spaCy [1] to extract all noun phrases from the narrations. Subsequently, we filter out noun phrases that describe the camera wearer and other people (*e.g.*, "*#C*", "*man X*", "*woman Y*", "*person*") or involve hand (*e.g.*, "*left hand*", "*right hand*", "*hands*") to get the object phrases for our grounding process.

Training details. We train the model using AdamW optimizer [11] with an initial learning rate of $2e^{-5}$, decayed with a cosine learning schedule. We freeze the CLIP image encoder and text encoder, and optimize the MLP layers in Region Encoder. The total number of parameters in our model is 150M and we only optimize 0.5% of them (787k). The batch size is 32, and the learnable temperature parameter τ is initialized with 0.07. For the Temporal Adaptive Pooling function, we set α_0 as 0.1 and β as 0.999. To accelerate training, we store mask proposals generated by SAM, and extract context-aware region embeddings for all mask proposals. During training, we load these pre-extracted region embeddings and optimize the MLP in region encoder. Training the model on eight Tesla V100 GPUs takes approximately 6 hours.

Inference time. Inference takes roughly 2.24 seconds per frame on a NVIDIA V100 GPU, with a significant portion of time ($2.1s$) being consumed by SAM mask generation.

Figure 1. 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⟩.

This is a common limitation in most object segmentation methods [9, 21, 22] that rely on mask proposals. The inference time can be significantly reduced through the adoption of more efficient mask proposal generators, such as FastSAM [24] and EfficientSAM [20]. For example, FastSAM can reduce the inference time to $0.39s$. However, it leads to a lower upper bound of $J\&F$ from 73.0% to 66.0%, illustrating the inherent trade-off between speed and accuracy.

Open-Vocabulary segmentation baselines. We use the official GitHub repositories to implement ODISE[1] [21] and GroundedSAM[2] [6, 10]. To adapt the open-vocabulary segmentation baselines to NVOS tasks, we do not pre-define a taxonomy of all object classes. Instead, we use the list of object phrases in the narration as the potential class names for each video clip to output segmentation masks of each object phrase.

2. VISOR-NVOS Benchmark Details

Annotation setup. The VISOR dataset [2] provides segmentation masks for annotated objects but lacks associated narrations, making it unsuitable for direct evaluation of NVOS methods. To address this limitation, we first extracted associated narrations from EPIC-Kitchens based on the timestamp of each frame. However, EPIC-Kitchens narrations are very short, typically consisting of only one verb and one noun (*e.g.*. "*stir pasta*", "*pour salt*", "*set down cutlery*"). These short-form narrations lack the richness required for NVOS task and fail to measure a method's capability to ground multiple objects. To mitigate this, we extended VISOR to VISOR-NVOS by collecting detailed, object-based narrations, as illustrated in Figure 1. The annotators are instructed to watch a 2-second video clip with the frame with annotated segmentation masks as the mid-

[1] https://github.com/NVlabs/ODISE
[2] https://github.com/IDEA-Research/Grounded-Segment-Anything

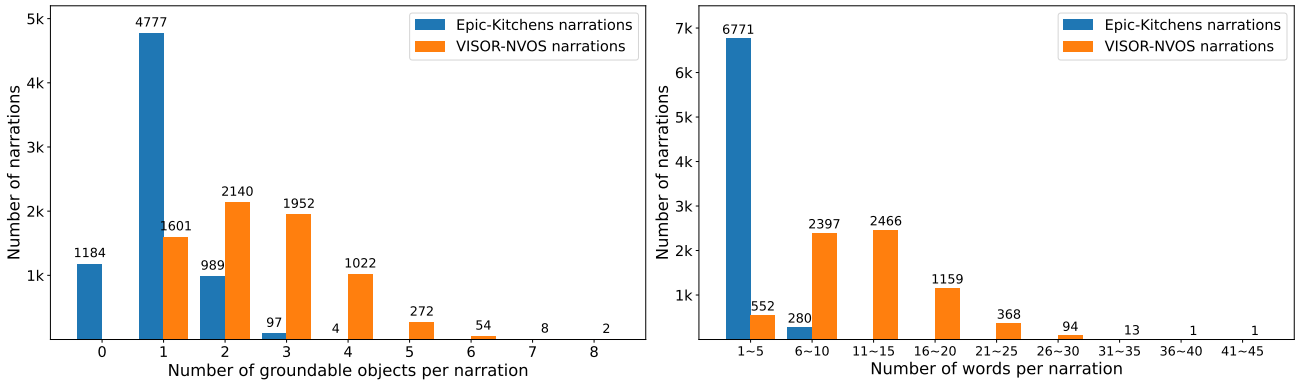


Figure 2. Dataset statistics for EPIC-Kitchens narrations and our annotated VISOR-NVOS narrations. *Left*: Histogram of the number of groundable objects per narration. *Right*: Histogram of the number of words per narration.

dle frame. Provided with a list of objects of interest and the original short EPIC-Kitchens narration, their task is to compose a comprehensive narration covering as many objects related to the user's activity in the video as possible, while disregarding irrelevant objects (*e.g.*, "*pesto sauce jar*" in the provided example). We ask the annotators to write the narration in a predefined format, *i.e.*, [referred object] followed by ⟨object ID⟩, as shown in Figure 1. After annotation, we parse these narrations to get the clean narration, and a list of object phrases as well as their associated object IDs.

Dataset statistics. As VISOR [2] does not release annotations for the test set, we use videos from VISOR validation set as the test set for VISOR-NVOS. Additionally, we annotate a subset of VISOR training videos to create our validation set. We have 7,561 video clips in the validation set and 7,051 video clips in the test set. Each video clip is annotated with one detailed narration. We use the validation set to select our best-performing model, and report the performance on the test set. In Figure 2, we present a histogram illustrating the number of groundable objects and the number of words per narration for narrations from Epic-Kitchens and our collected narrations on VISOR-NVOS test set. Notably, the majority of narrations from EPIC-Kitchens are very short, ranging from one to five words, and including only one groundable object. Particularly, 1,184 EPIC-Kitchens narrations have no groundable object. In contrast, our annotated narrations on VISOR-NVOS are longer and contain multiple groundable objects, making it more suitable for NVOS evaluation.

Evaluation setup. VISOR contains both sparse annotation (manually segmented masks) and dense annotation (dense masks obtained automatically through interpolation). Following the established setup for semi-supervised VOS [3], we evaluate only on sparsely annotated segmentation masks. The dense masks remain a valuable resource for evaluating the model's tracking capability and consistency across multiple consecutive frames, which we leave as a future direction. We pre-train our model on Ego4D, and evaluate our model on VISOR-NVOS benchmark without fine-tuning. It is important to note that this benchmark evaluates the zero-shot transfer capability of our approach, *i.e.*, our model has not been trained on any videos from EPIC-Kitchens or any narrations from VISOR-VNOS. In addition, while we use these annotated narrations to benchmark our NVOS task, our annotations can be used for other related tasks, such as video captioning and video-text retrieval.

Comparison with other datasets. We compare our VISOR-NVOS benchmark with other related datasets in Table 1. Three key characteristics set our VISOR-NVOS benchmark apart from previous datasets: 1) it focuses on egocentric videos rather than third-person view videos, 2) annotations are provided in the form of segmentation masks instead of bounding boxes, and 3) each video is associated with a narration that describes the context and contains multiple groundable objects. While our benchmark shares a similar scale with UVO-VNG [18], it features more groundable objects in each narration and focuses on egocentric videos rather than third-person view videos. In comparison to YouCook2-BB [26], a previous dataset in video object grounding, our dataset excels in several aspects: it includes a larger number of video narrations, a greater variety of objects, finer-grained segmentation masks instead of bounding boxes, and a distinctive emphasis on egocentric videos.

3. Details on Other Evaluation Datasets

VOST. We evaluate our model on the validation set of VOST. VOST contains videos from both EPIC-Kitchens and Ego4D, and we remove the videos existing in our training set from the evaluation set. In total, we have segmentation masks for 7,820 frames from 70 videos, including 50 cooking-related videos and 20 non-cooking videos. Each

Dataset	Source	Task	Type	# Vid.	# Narr.	# Obj. (per narr.)	ego
RefEgo [8]	Ego4D [5]	Referring Expression Loc.	bbox	12,038	12,038	12,038 (1)	Yes
ANet-Entities [28]	ActivityNet [7]	Video Object Localization	bbox	14.9k	51.8k	157.8k (3.05)	No
YouCook2-BB [26]	YouCook2 [27]	Video Object Grounding	bbox	647	4,325	9,766 (2.26)	No
EgoHOS [23]	multiple †	Hand Object Segmentation	mask	11,243	N/A	17,568 (N/A)	Yes
OVIS-VNG [18]	OVIS [12]	Video Narrative Grounding	mask	505	1,554	2,407 (1.55)	No
UVO-VNG [18]	UVO [19]	Video Narrative Grounding	mask	7,587	22,749	43,058 (1.89)	Mixed
VOST * [17]	EPIC-Kitchens+Ego4D	Narration-based VOS	mask	70	7,820	7,820 (1)	Yes
VISOR-NVOS	EPIC-Kitchens [2]	Narration-based VOS	mask	14,612	14,612	37,170 (2.54)	Yes

Table 1. Comparison of VISOR-NVOS with existing related datasets. * The subset of VOST videos used as our evaluation benchmark. † EgoHOS contains videos sourced from Ego4D [5], EPIC-Kitchens [2], THU-READ [16], and self-collected GoPro videos.

frame is annotated with the segmentation masks of one or multiple object instances of the same object class. The average number of object instances per frame is 1.54.

YouCook2-BB. We follow the setup in NAFAE [13] to conduct bounding box evaluation and the setup in CoMMa [15] to conduct point prediction evaluation. CoMMa predicts an attention map for each object, and if the highest attention similarity score lies in the ground-truth bounding box, the result counts as a "hit" and otherwise it counts as as a "miss". The point accuracy is calculated as a ratio between hits to the total number of predictions $\frac{\#\text{hits}}{\#\text{hits}+\#\text{misses}}$.

4. Additional Experiments

4.1. Ablation Studies on Dual Encoder

Context-Aware Region Encoder. We investigate the impact of our Context-Aware Region Encoder in Table 2a. We report the performance of the models in two scenarios: without learning (*i.e.*, we replace the MLP of our region encoder with the identity mapping), and with learning on Ego4D using our proposed Global-Local Contrastive Learning objectives. We first replace the CLIP image encoder with a same-sized ViT model pre-trained on ImageNet-21k [4] to assess the advantages derived from extensive vision-language pre-training. Directly employing the ImageNet pre-trained model for grounding, without additional learning, results in poor performance due to a misalignment between the region and phrase embeddings. After learning on Ego4D, we observe a significant improvement in $\mathcal{J}\&\mathcal{F}$ to 21.5%, which demonstrates that our proposed framework is able to learn an improved alignment for region-phrase pairs even without a reasonable initialization. However, a significant performance gap of 16.6% persists when compared to ROSA, showing that our proposed Context-Aware Region Encoder is able to effectively leverage the advantages of large-scale vision-language pre-training from CLIP. In addition, we compare our Region Encoder with Crop+Mask and MaskCLIP [25]. Our Region Encoder is not only more efficient than cropping and masking, but also improves

Region Encoder	w/o learning	w/ learning
ImageNet Pretrained	3.4	21.5
Crop+Mask	24.3	-
MaskCLIP [25]	28.1	36.9
ROSA	**28.1**	**38.1**

(a) Performance of $\mathcal{J}\&\mathcal{F}$ of various region encoders without and with learning on Ego4D.

Phrase Encoder	All	Exhaust.	Inexhaust.
narration-aware	32.5	34.7	17.0
localized	37.0	39.3	19.4
average	**38.1**	**40.5**	**21.5**

(b) Effect of context in phrase encoder.

Phrase Encoder	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
no MLP	**34.9**	**41.2**	**38.1**
add 1 layer	28.3	34.9	31.6
add 2 layers	27.4	33.8	30.6

(c) Effect of MLP in phrase encoder.

Table 2. Ablations on CLIP-based Dual Encoder on VISOR-NVOS test split.

the $\mathcal{J}\&\mathcal{F}$ by 3.8%, showing the advantages of using contexts for grounding task. Due to the computation cost of Crop+Mask, we are unable to train the model on Ego4D. In comparison to MaskCLIP, ROSA exhibits comparable performance without learning, and outperforms MaskCLIP by 1.2% after learning using our proposed objectives.

How does the context in phrase encoder help grounding? The objects on VISOR are annotated with an "exhaustive" label, indicating whether the object has been exhaustively annotated, *i.e.*, there are no more instances of the same class in the view. We evaluate the performance of the narration-aware and localized phrase embeddings for exhaustive and inexhaustive objects in Table 2b. Merely

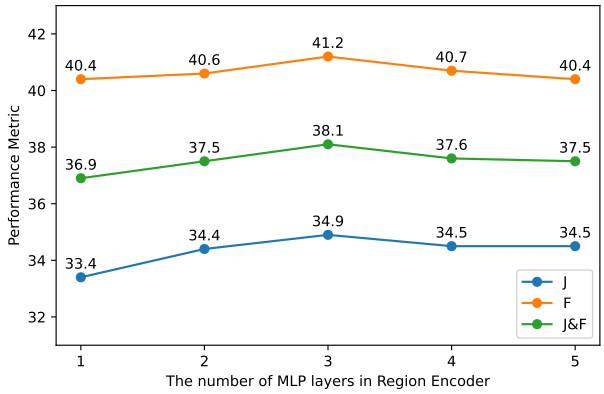


Figure 3. Performance w.r.t. the number of MLP layers in Region Encoder (VISOR-NVOS test split).

using narration-aware phrase embeddings leads to suboptimal performance, as the contextualized embeddings from CLIP's text encoder do not well emphasize the local information. However, compared with only using localized phrase embeddings, the combination of narration-aware and localized embeddings demonstrates an overall improvement in $\mathcal{J}\&\mathcal{F}$ by 1.1%. In particular, adding narration-aware phrase embeddings leads to 2.1% improvement for inexhaustive objects, which shows the importance of context for inexhaustive objects.

MLP in Region Encoder and Phrase Encoder. Figure 3 illustrates the performance variations as we change the number of MLP layers in Region Encoder. The model achieves the best performance with three layers. We also explored adding MLP in Phrase Encoder, which worsens the performance, as detailed in Table 2c. We postulate that freezing the CLIP text encoder is beneficial in preserving the model's generalization capability.

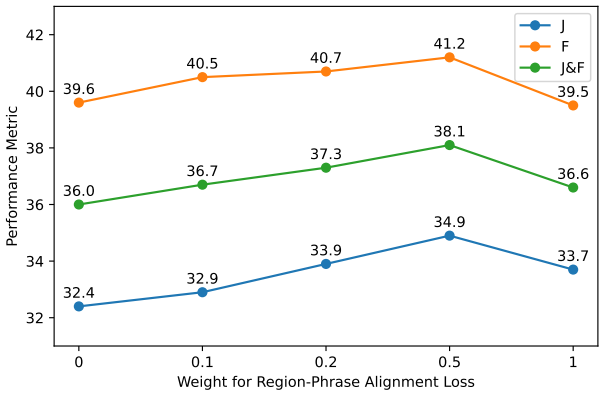
What is the effect of ViT model size? In Table 3, we report the performance when we use CLIP models of various sizes before and after learning on Ego4D. Our learning approach is able to achieve improvement in all cases. We notice a significant performance gain from ViT-B/32 to ViT-B/16, which implies the importance of higher-resolution maps for achieving fine-grained egocentric video understanding.

4.2. Additional Ablations on Global-Local Contrastive Learning

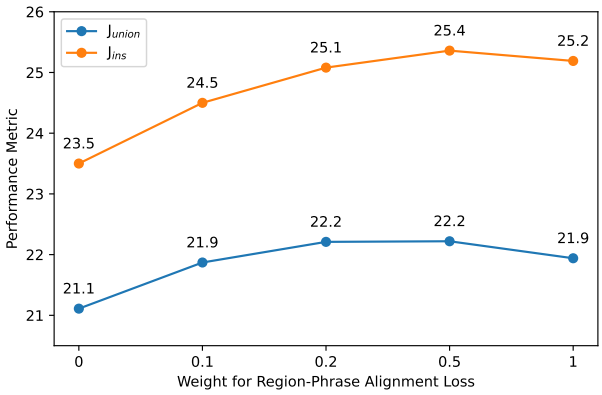
What is the effect of RPA loss weight? Figure 4 shows the performance of ROSA under varying weights for the proposed region-phrase alignment loss on VISOR-NVOS and VOST. When the weight λ is 0, our model relies solely on global video-narration contrastive loss to learn region-phrase alignments. When λ is larger than 0, we observe a consistent improvement in all metrics, showing the effectiveness of the proposed local region-phrase contrastive loss. The model performs best when λ is 0.5.

Model Size	VISOR-NVOS			VOST	
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}_{union}$	$\mathcal{J}_{ins}$
w/o learning					
ViT-B/32	17.9	25.1	21.5	10.6	11.9
ViT-B/16	24.8	31.3	28.1	16.3	19.7
ViT-L/14	26.6	32.9	29.8	17.8	20.6
w/ learning					
ViT-B/32	22.6	30.0	26.3	11.4	12.8
ViT-B/16	34.9	41.2	38.1	22.2	25.4
ViT-L/14	38.7	46.0	42.4	23.2	26.7

Table 3. Performance of different model sizes.



(a) VISOR-NVOS



(b) VOST

Figure 4. Performance w.r.t. different weights for region-phrase alignment loss.

How does the model perform on unseen/rare object phrases? We categorize the object phrases based on their occurrences in the training set and evaluate their perfor-

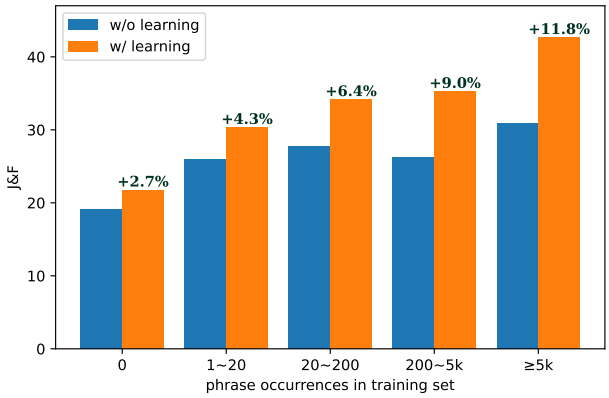


Figure 5. Performance gain on VISOR-NVOS test split from weakly-supervised training on Ego4D using our proposed Global-Local Contrastive Learning framework w.r.t. phrase occurrences in the training set.

Method	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
w/o learning	24.8	31.3	28.1
fix alignments	29.0	35.7	32.4
update alignments (ours)	**34.9**	**41.2**	**38.1**

Table 4. Comparison on using fixed alignments and updated alignments during training.

mance in Figure 5. The figure illustrates the performance gain achieved through our Global-Local Contrastive Learning framework with respect to the phrase occurrences in the training set. For phrases comprising multiple words, we consider one occurrence whenever any word from the phrase appears in the narration. Overall, our contrastive learning exhibits greater improvements for object phrases with higher occurrences in the training set. There is a noteworthy 2.7% improvement for phrases that never occur in the training set, showcasing the model's ability to generalize to unseen data. Moreover, the model demonstrates a substantial 4.3% improvement for phrases with very few occurrences (1 to 20) in the training set, highlighting the data efficiency of our learning approach.

How does learning improve region-phrase alignments? In our implementation, region-phrase alignments are determined based on the similarity between the learned region embeddings and phrase embeddings. Additionally, we investigate a scenario where the alignments are fixed using the initial region embeddings (without MLP) throughout the training process. In other words, we generate pseudo labels using the initial alignments and fix them during training. As outlined in Table 4, using fixed alignments also improves $\mathcal{J}\&\mathcal{F}$ by 4.3% after learning. However, dynamically updating alignments surpasses it by 5.7%.

Method	All	Cooking	Non-Cooking
SAM+CLIP	16.5|19.2	17.9|21.7	15.0|15.3
ROSA	**22.2|25.4**	**24.5|29.6**	**20.0|20.0**

Table 5. Comparison between SAM+CLIP and ROSA on VOST in cooking videos and non-cooking videos ($\mathcal{J}_{union}|\mathcal{J}_{ins}$).

4.3. Other Analysis

How does the model generalize to non-cooking videos? While both our training set and VISOR-NVOS focus on cooking domain, the VOST dataset contains both cooking-related and non-cooking (*e.g.*, "*mold clay*", "*cut paper*") videos. We use this dataset to evaluate the domain generalization capability of our model in Table 5. Despite being pre-trained on cooking-related videos, our model improves the performance of SAM+CLIP in both cooking and non-cooking videos. The improvement of 5.0% in $\mathcal{J}_{union}$ and 4.7% in $\mathcal{J}_{ins}$ in non-cooking videos, while smaller than that in cooking videos, underscores the generality of our model across varied video contexts.

How does the model perform on different objects? The detailed performance breakdown for each object phrase is illustrated in Figure 6. To ensure statistical significance and reduce the impact of randomness, we exclusively present results for object phrases occurring more than 50 times in the test set. The model performs well on objects like "chopping board", "bowl", and "table". Interestingly, it struggles with objects related to "tap water", "tap", and "water", which can be ambiguous or transparent. Additionally, it encounters challenges with objects that frequently appear in multiple instances within a scene, such as "sponge" and "fork".

5. Limitations and Ethical Concerns

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. While we show that our weakly-supervised training also improves grounding on non-cooking videos in Table 5, a performance gap still exists between the two domains. Additionally, our model's performance is limited by the quality of mask proposals generated by SAM. While there is room for improvement between the best performance (42.7%) and the upper bound of SAM proposals (73.0%), any future improvements will be constrained by the upper bound without further advancements on mask

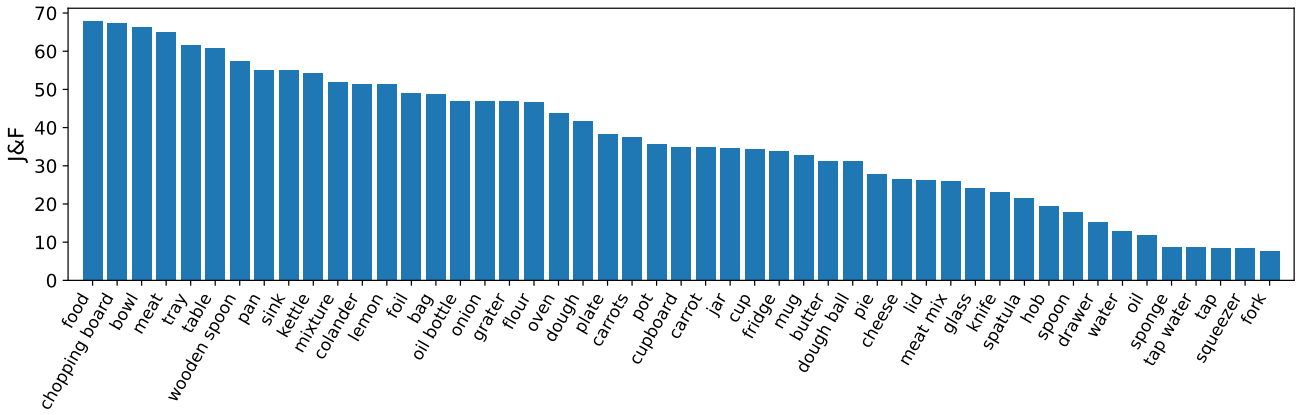


Figure 6. Breakdown performance per object phrase on VISOR-NVOS test split. To ensure statistical significance and reduce the impact of randomness, we exclusively present results for object phrases occurring more than 50 times in the test set.

proposal generation methods. Regarding ethical concerns, we use the videos and narrations from the public egocentric video dataset Ego4D [14], which may have gender, age, geographical and cultural bias.

References

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

[2] 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, 2, 3

[3] 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

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[5] 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. 3

[6] 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. 1

[7] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 3

[8] 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. 3

[9] 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. 1

[10] 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. 1

[11] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 1

[12] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision*, 130(8): 2022–2039, 2022. 3

[13] 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. 3

[14] 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. 6

[15] 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. 3

[16] Yansong Tang, Yi Tian, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Action recognition in rgb-d egocentric videos. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3410–3414. IEEE, 2017. 3

[17] 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. 3

[18] 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, 3

[19] Weiyao Wang, Matt Feiszli, Heng Wang, and Du Tran. Unidentified video objects: A benchmark for dense, open-world segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10776–10785, 2021. 3

[20] Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, et al. Efficientsam: Leveraged masked image pretraining for efficient segment anything. *arXiv preprint arXiv:2312.00863*, 2023. 1

[21] 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. 1

[22] 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. 1

[23] Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *European Conference on Computer Vision*, pages 127–145. Springer, 2022. 3

[24] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything. 2023. 1

[25] 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

[26] Luowei Zhou, Nathan Louis, and Jason J Corso. Weakly-supervised video object grounding from text by loss weighting and object interaction. *BMVC*, 2018. 2, 3

[27] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 3

[28] 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. 3