SLVP: Self-supervised Language-Video Pre-training for Referring Video Object Segmentation

Jie Mei[1]*, AJ Piergiovanni[2], Jenq-Neng Hwang[1], Wei Li[2]
[1]University of Washington, Seattle jiemei@uw.edu, hwang@uw.edu
[2]Google Research, Brain Team ajpiergi@google.com, mweili@google.com

Abstract

The referring video object segmentation (R-VOS) task requires a model to understand both referring expression and video input. Most recent works are mainly based on an encoder-decoder type of architecture. Although their text and visual encoders can benefit from separately pre-trained backbones, their decoder is trained from scratch on a combination of image/video segmentation datasets. However, pixel-wise annotation with referring expressions is extremely expensive which makes it challenging to further improve the performance. Due to the same reason, current vision-language pre-training works mainly focus on learning general feature representations for image-level or object-level tasks, which may be not optimal for the downstream pixel-level segmentation task. To bridge this gap, we present a general self-supervised language-video pre-training (SLVP) architecture. With the relatively cheap video caption dataset, SLVP can learn pixel-level features by introducing optical flow as the intermediate target during pre-training. Correspondingly, we propose simple transfer learning models that can reuse pre-trained modules for the downstream R-VOS task. Furthermore, the proposed general SLVP architecture can support either 'language as query' fusion or 'vision as query' fusion. Experiments show the superiority of the under-studied 'vision as query' method which can achieve better performance than the state-of-the-art methods on Ref-Davis17 and Ref-Youtube-VOS benchmarks even with fewer model parameters. We further adopt the challenging VISOR benchmark to the R-VOS task and our SLVP serves as the first strong baseline for R-VOS task on it.

1. Introduction

Referring video object segmentation (R-VOS) is an emerging multi-modal task, requiring the model to segment the specific object referred by a language description in all input frames. This task is gathering great attention in the research community because of the potential benefits to many applications in an interactive way, e.g., video editing and video surveillance. Compared with the traditional video-object segmentation (Semi-VOS) task [31, 43] which assumes the availability of ground-truth mask annotation in the first frame during inference, the R-VOS task is more challenging because it requires the model to have a comprehensive understanding of the raw input videos and language description without any available mask during inference. Therefore, the model should know what the target object is described by the referring expression and then accurately segment it from the raw video.

Existing approaches for the R-VOS task can be categorized into three groups: (1) Bottom-up approaches. These approaches directly decode the target object masks using fully convolution networks (FCNs) [21] based on vision-language fused features. (2) Top-down approaches. These approaches first segment all potential objects in each frame using an instance segmentation model then associate each object using a tracking algorithm. Finally, the target object masks are selected based on the language description. (3) Language as queries approaches. These methods are an encoder-decoder type of architecture, which takes advantage of the query mechanism in Transformer [38], treating referring expressions as queries and still using some convolution heads to decode the object mask.

These three streams of approaches have shown promising results but share an intrinsic limitation, i.e., only some parts of the model can benefit from pre-training such as backbones while the remaining parts of the model can only be trained from scratch on a combination of image/video referring segmentation datasets. This makes it challenging to further improve the model performance since pixel-level annotations are extremely expensive. Besides, current pre-training strategies are mainly designed for image-level or object-level tasks. For example, existing vision-language pre-training strategies can utilize a large amount of relatively cheap image-text pairs [33] or object bounding-box-text pairs [17] and inherently benefit down-

*Work done during an internship at Google Brain.

stream image-level or object-level tasks. On the other hand, self-supervised pre-training strategies may light the way to help pixel-level tasks since they show vision transformers such as [2] contain explicit information about the semantic segmentation of an image even when there are no labels during pre-training. However, most existing works focus on single-modality self-supervised pre-training such as DINO [2] and MAE [10].

Thus a natural question is when there are no pixel-level annotation datasets available, how to design a multi-modal *i.e.,* language and video, self-supervised **pre-training strategy** to learn pixel-level semantic information. Furthermore, the second question is how to design the **model architecture** so that the pre-training strategy can benefit the whole model and further bring improvement to the downstream pixel-level and temporal-based R-VOS task. The above questions motivate us to design a synchronous pre-training and transfer-learning architecture to tackle the R-VOS task elegantly. In contrast to existing approaches, our decoder served for the fusion purpose can also benefit from the pre-training. Thanks to its simplicity, our general architecture supports not only the **'language as query'** fusion method but also the under-studied **'vision as query'** fusion method to be explained in Sec. 3.

The main contributions of this work are as follows. (1) We propose a self-supervised language-video pre-training strategy that can leverage relatively cheap video-caption datasets to make the decoder learn temporal semantic information based on video and text input. Experiments show the self-supervised pre-trained decoder can bring non-negligible improvement to the downstream R-VOS task. (2) We present a synchronous transfer learning architecture for the R-VOS task that can maximumly benefit from the pre-trained model. It shares modules as much as possible with the pre-training architecture and employs a simple shared linear mask head on each token. (3) Experiments show the superiority of the under-studied 'vision as query' method and that even when there are fewer segmentation training data or fewer model parameters, our proposed method can achieve on-par or even better performance than the state-of-the-art methods.

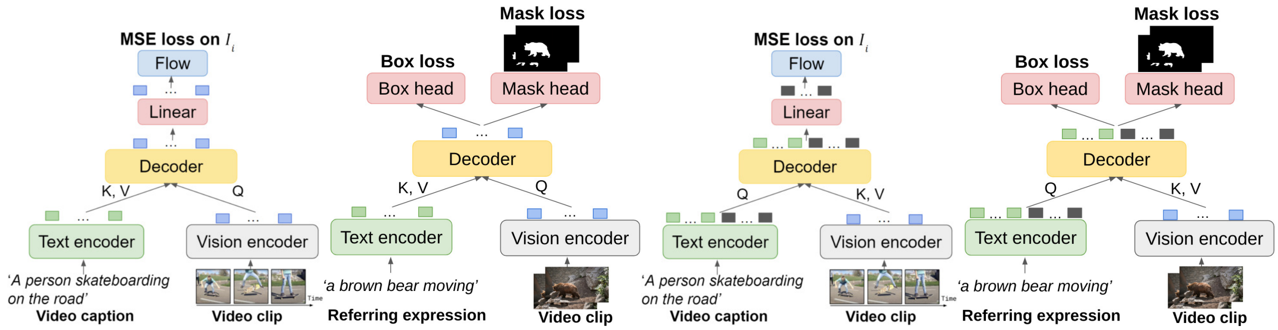
2. Related Work

Semi-supervised Video Object Segmentation. This related task assumes the ground-truth masks of target objects are available in the first frame during inference. Thus the model only needs to propagate these masks to other frames. Tracking the object based on feature matching is one mainstream approach in most recent works [4, 29, 40, 44]. STM [29] stores a memory of objects' features in the past frames and utilizes the attention mechanism to perform feature matching to predict the masks in the current frame. This single-modality-based task does not require the model to understand any language description.

Referring Video Object Segmentation. Referring video object segmentation (R-VOS) is a multi-modality task. It provides the language description instead of the first frame's mask ground truth for the target object during inference. Thus, it is a more challenging task. As mentioned previously, the current methods for R-VOS mainly follow three groups: (1) Bottom-up methods, which directly apply the image-based methods to each video frame independently [9, 14, 22, 49] without learning any temporal information to predict consistent masks. (2)Top-down methods, which first find many potential object tracklets using a tracking algorithm, and the target object is filtered out using a language grounding model [19] without considering the model complexity and heavy computation. (3) Language as query methods. The typical language as query methods, Referformer [41] and R^2-VOS [18], propose a transformer-based [38] encoder-decoder architecture to fuse language and vision features and apply dynamic convolution operation to decode masks for each target object. Although their text encoder and vision encoders are pre-trained on non-segmentation datasets, their decoders do not benefit from any pre-training. This makes it challenging to improve the performance since pixel-level annotation with referring expressions is extremely expensive.

In contrast to the above approaches, we propose a synchronous pre-training and transfer-learning architecture to tackle the R-VOS task elegantly. The proposed self-supervised pre-training model shares a similar architecture with the transfer-learning pipeline. Thus, our decoder can benefit from the pre-training on relatively cheap video-caption datasets. Thanks to its simplicity, *i.e.,* applying a shared linear mask head on each token, our architecture supports two fusion methods, i.e., 'language as query' or 'vision as query'. For both methods, we explore the gains from the proposed self-supervised language-video pre-training strategy.

Self-Supervised Learning Supervised training demonstrates outstanding performance in many tasks [13, 23–26, 26, 48, 50]. But with Transformers [38] successfully becoming a general building block in both language and vision, the computer vision community starts to bring in self-supervised representation learning methods such as MAE [10, 11] by referring to denoising/masked autoencoding methodology [39] introduced in BERT [8]. The features from the pre-trained model can achieve outstanding performance in image-level tasks such as zero-shot image classification. Recently, contrastive learning [3, 12, 30, 42] which models image similarity and dissimilarity between augmented views is getting popular. Besides single-modality training, existing vision-language pre-training strategies can utilize a large amount of relatively cheap image-text pairs [33] or object bounding-box-text pairs [17] and inher-



(a) Pre-training (b) Transfer Learning (c) Pre-training (d) Transfer Learning

Figure 1. The general SLVP architecture consists of four key components: text encoder, vision encoder, decoder served for the fusion purpose, and some linear heads for either intermediate optical flow prediction in the pre-training stage or bounding box and mask prediction in the transfer-learning stage. (a) (b) **Vision as Query** fusion method. (c) (d) **Language as Query** fusion method.

ently benefit downstream image-level or object-level tasks.

However, these approaches mainly focus on learning a good encoder. Although MAE [11] uses a decoder during the pre-training to reconstruct the original frame, the decoder is discarded for the downstream tasks. Our proposed self-supervised language-video pre-training strategy aims to also make the decoder learn temporal semantic information based on video and text input by leveraging relatively cheap video-caption datasets so that it can bring improvement to the downstream pixel-level R-VOS task.

3. Methodology

Given a video clip $\mathcal{V} = \{v_t\}_{t=1}^T$ with T frames and one corresponding referring expression $\mathcal{R} = \{r_l\}_{l=1}^L$ with L words, the R-VOS model is expected to produce T-frame binary segmentation masks for the referred object, *i.e.,* $\mathcal{M} = \{m_t\}_{t=1}^T, m_t \in \mathbb{R}^{H\times W}$, where H and W are the frame height and width. Our proposed general architecture, called self-supervised language-video pre-training (SLVP), is an encoder-decoder architecture based on pure transformer modules. By applying a shared linear head on each token to get the prediction, the proposed general architecture can support not only 'language as query' fusion but also 'vision as query' fusion. Furthermore, due to the consistency between the proposed pre-training and transfer-learning architectures, our decoder served for the fusion purpose can benefit from pre-training and contribute a non-negligible improvement to the downstream R-VOS task. Details of using SLVP in pre-training and transfer-learning for 'vision as query' and 'language as query' fusion methods are in Fig. 1.

3.1. Vision as Query

Existing methods mainly use 'language as query' fusion. For example, Referformer [41] introduces some learnable queries conditioned on text features as input to the decoder, adopted from Deformable-DETR [1, 51] while our general SLVP architecture supports under-studied 'vision as query' fusion as shown in Fig. 1a and 1b.

Self-supervised Pre-training. We use a relatively cheap non-segmentation dataset for pre-training, *i.e.,* a video-caption dataset. The intermediate pre-training target is to predict the optical flow, $o_{i,1} \in \mathbb{R}^{H\times W\times 2}$, between the first frame and any ith frame based on the caption and video input without any optical flow ground truth. Thus, the self-supervised loss function is applied on the original ith frame and an RGB image, I_i, reconstructed by applying $o_{i,1}$ on the original first frame:

$$\begin{aligned} I_i &= Flow(o_{i,1}, V_1), \\ L_{MSE} &= |I_i - V_i|^2, \end{aligned} \tag{1}$$

where the mean square error is used as the loss function, $Flow(\cdot)$ operation is denoted as the blue module in Fig. 1a and there are no learnable parameters because this operation just moves every pixel in the first frame V_1 to a new location based on $o_{i,1}$.

To predict the optical flow $o_{i,1} \in \mathbb{R}^{H\times W}$ for the ith frame from a sequence of text features $f_r \in \mathbb{R}^{L\times d}$ output from the text encoder and a sequence of frames' features $f_v \in \mathbb{R}^{(T\times N)\times d}$ (N is the number of tokens of one frame, *i.e.,* the number of patches of one frame), we first use a Transformer that consists of self-, cross-attentions and feed-forward network as our decoder to fuse the text features and frames feature by using frames' features as query and text features as key and value:

$$f_{fused} = \mathbb{T}(f_v + E_{pos} + E_{tem}, f_r), \tag{2}$$

where $\mathbb{T}$ is a Transformer, $E_{pos} \in \mathbb{R}^{N\times d}$ and $E_{tem} \in \mathbb{R}^{T\times d}$ are the learnable positional and temporal encodings respectively, the first input of $\mathbb{T}$ is the query and the second is key and value. E_{pos} and E_{tem} will be repeated by T, N times respectively when added to f_v. The output $f_{fused} \in \mathbb{R}^{(T\times N)\times d}$ is the fused sequence of frames' features.

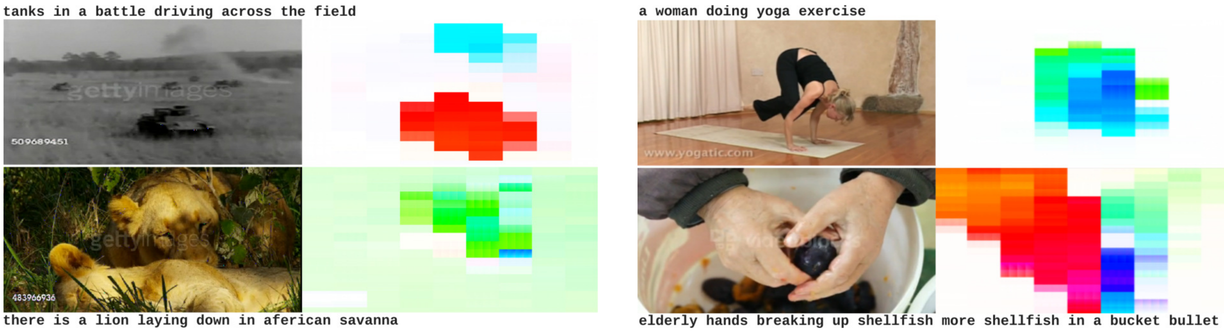


Figure 2. **Self-supervised learned optical flow** in pre-training stage. Each image is a sampled frame of its own video from (S-MiT) dataset [28]. The corresponding video caption is on the top or bottom of each frame. On the right side of each frame, there is the self-supervised learned optical flow between the shown frame and the first frame in its own video. Predicted directions and the extent of pixel movement are visualized with different colors.

Finally, inspired by MAE [11] ,we use a shared linear layer on each token of $f_{fused} \in \mathbb{R}^{(T\times N)\times d}$ to transform the last dimension from d to $p \times p \times 2$, where p is the patch size and 2 represents two-dimensional optical flow values for each pixel:

$$\hat{f}_{fused} = Linear(f_{fused}), \tag{3}$$

where $\hat{f}_{fused} \in \mathbb{R}^{(T\times N)\times(p\times p\times 2)}$ is the generated optical flow for each image patch. Then we reshape $\hat{f}_{fused}$ into $\mathbb{R}^{T\times H\times W\times 2}$ to get optical flow, $o_{i,1} \in \mathbb{R}^{H\times W\times 2}$, for the ith frame. A predicted optical flow demo is shown in Fig. 2.

In detail, our vision encoder is applied on each frame independently while the decoder takes in the concatenated sequence of frames' features. This is because it is necessary to allow the decoder to observe nearby frames before it can predict the meaningful optical flow. Thus, the proposed architecture is a temporal-based method. When only observing the input and output of the proposed pre-training architecture, we can see it predicts optical flow for each input image patch.

Transfer Learning. The transfer learning architecture share all the modules with the pre-training architecture except for the last linear layers. Thus all encoders, decoders, and learnable positional/temporal encodings are initialized with the pre-trained weights.

Since the last shared linear layer in the pre-training is trained to predict optical flow, we replace it with another two reinitialized linear heads, *i.e.,* one for bounding box regression and the other for binary mask prediction for the downstream R-VOS task as shown in Fig. 1b. The box linear head is applied on the max-pooled fused features, $f_{max} \in \mathbb{R}^{T\times d}$, and the mask linear head is shared among all tokens:

$$\begin{aligned} f_{max} &= MaxPooling(f_{fused}), \\ B &= Linear_{box}(f_{max}), \\ M &= \sigma(Linear_{mask}(f_{fused})), \end{aligned} \tag{4}$$

where $B \in \mathbb{R}^{T\times 4}$ is the predicted bounding box for T frames, $\sigma(\cdot)$ is the sigmoid operation, $M \in \mathbb{R}^{(T\times N)\times(p\times p\times 1)}$ is the predicted binary mask for each image patch. M will be reshaped into $\mathbb{R}^{T\times H\times W}$ and then we can get a binary mask, $m_i \in \mathbb{R}^{H\times W}$, for the ith frame.

When only observing the input and output of the proposed transfer learning architecture, we can see it predicts binary masks for each input image patch based on temporal information and referring expression. It also outputs the regressed bounding boxes for each frame.

For box loss, we use GIoU [35] and L1 loss; for mask loss, we use Dice loss [37] and binary cross-entropy.

Box loss. If we denote a predicted bounding box as $B_p(x_1, y_1, x_2, y_2)$ and the ground truth bounding box as $B_g(X_1, Y_1, X_2, Y_2)$, then GIoU [35] is defined as following:

$$\begin{aligned} IoU &= \frac{|B_p \cap B_g|}{|B_p \cup B_g|}, \\ GIoU &= IoU - \frac{|C \backslash (B_p \cup B_g)|}{|C|}, \end{aligned} \tag{5}$$

where C is the smallest enclosing bounding box for B_p and B_g; the nominator in the second equation is the area occupied by C excluding B_p and B_g. IoU has a major weakness when used as a loss function: if $IoU(B_p, B_g) = 0$, IoU can not reflect if two bounding boxes are in the vicinity of each other or very far from each other. However, $GIoU$ takes the smallest enclosing bounding box into consideration to overcome this issue. Finally, the $GIoU$ loss is $1 - GIoU$.

L1 loss is a straightforward loss between four coordinates (top-left point and bottom-right point) of B_p and B_g:

$$L_1 = |x_1 - X_1| + |y_1 - Y_1| + |x_2 - X_2| + |y_2 - Y_2|. \tag{6}$$

Mask loss. Dice loss [37] and binary cross entropy (BCE) loss are as follows:

$$Dice\ Loss = 1 - \frac{2\sum_i^N m_i g_i}{\sum_i^N m_i^2 + \sum_i^N g_i^2},$$
$$BCE\ Loss = \frac{1}{N}\sum_{i=1}^{N} -\left(g_i \log\left(m_i\right) + \left(1-g_i\right)\log\left(1-m_i\right)\right), \tag{7}$$

where N is the total number of pixels; m_i and g_i are the values in the predicted mask M and ground-truth binary mask G respectively. Finally, the total loss is the summation of Dice loss, BCE loss, GIoU loss, and L1 loss. The coefficients for losses are set as $\lambda_{L1} = 5$, $\lambda_{dice} = 5$, $\lambda_{giou} = 1$, and $\lambda_{bce} = 1$.

3.2. Language as Query

Our general SLVP architecture also supports the 'language as query' fusion method with a slight modification as shown in Fig. 1c and 1d.

Self-supervised Pre-training. Since we still hold the same spirit mentioned in the pre-training, *i.e.,* predicting the optical flow for each image patch, we have to make sure the length of tokens output from the decoder is the same as the input sequence of frames' features $f_v \in \mathbb{R}^{(T\times N)\times d}$. Thus, we create a shared learnable query token, $q \in \mathbb{R}^d$, and repeat it by $T \times N$ times to get $\hat{q} \in \mathbb{R}^{(T\times N)\times d}$, denoted as gray cubes in Fig. 1c. Then we fuse the text features and frames features by using text's features $f_r \in \mathbb{R}^{L\times d}$ concatenated by $\hat{q}$ as query and frame features as key and value:

$$f_{fused} = \mathbb{T}(cat(f_r, \hat{q} + E_{pos} + E_{tem}), f_v + E_{pos} + E_{tem}), \tag{8}$$

where output $f_{fused} \in \mathbb{R}^{(L+(T\times N))\times d}$ is the fused sequence of features. But we only use the last $T \times N$ tokens as the input to the later shared linear layer to predict the optical flow for each image patch. The other parts including the loss function in the architecture are the same as those of the 'vision as query' architecture.

Transfer Learning. Same as the 'vision as query' architecture, we also replace the last shared linear layer in the pre-training with another two reinitialized linear heads for bounding box regression and binary mask prediction respectively as shown in Fig. 1d. In both the 'vision as query' and 'language as query' methods, our decoder served for the fusion purpose can benefit from the self-supervised pre-training.

4. Experiments

4.1. Implementation Details

Model Settings We use T5-pretrained text encoder [34], and CoCa-pretrained visual encoder [46], denoted as 'Pre-trained Es' in all experiment tables. Each of the encoders has 12 transformer self-attention layers. We use an 8-layer transformer, that consists of self-, cross-attention, and feed-forward networks, as our decoder. For both pre-training and transfer learning, we use 18 as patch size, 360×648 as the frame resolution, 64 as the maximum sentence length, and 4 as video clip length.

Pre-training Details During our pre-training, we freeze the text and vision encoders. This is because we want to see the improvement contributed only by the self-supervised pre-trained decoder on the downstream pixel-level R-VOS task. Besides, we use the sliding windows to obtain the short clips from videos and each clip consists of 4 randomly sampled frames with 6 as the sampling rate to cover enough object movement. There is no augmentations used during pre-training.

Transfer Learning Details In both 'vision as query' and 'language as query' methods, we concatenate the bounding box prediction with each of the fused tokens before applying the mask linear head so that the mask prediction can consider the object location. We also use random-flip, random-crop augmentation, and color-jittering during transfer learning, denoted as 'Augs', in all experiment tables. During the inference on R-VOS benchmarks, we directly output the predicted segmentation masks without any post-processing such as mask propagation [44] used in some previous works so that we can see the authentic segmentation improvement contributed by the pre-trained decoder.

4.2. Datasets and Metrics

Pre-training Dataset We use the large-scale Spoken Moments in Time (S-MiT) dataset [28] as the pre-training dataset. It consists of 500K pairs of video clips and corresponding captions depicting a broad range of different dynamic events. The captions are semantically rich compared to simple action labels. S-MiT covers a subset of the videos in the Moments in Time dataset [27]. The clips are 3 seconds long. On average, the captions have a length of 18 words and contain 1.58 verbs. Thus these attributes make it well-suited for our self-supervised pre-training target, *i.e.,* predicting the optical flow for frames.

R-VOS Benchmarks After the pre-training, we finetune and evaluate the models on Ref-Davis17 [16] and Ref-Youtube-VOS [36]. **Ref-Youtube-VOS** [36] is a large-scale benchmark that covers 3,978 videos with about 15K language descriptions. Among them, 3,471 videos are for training and 202 videos are for validation. For a fair comparison, we follow ReferFormer's [41] training setup, i.e., before finetuning on Ref-Youtube-VOS, we also first finetune our pre-trained model on RefCOCO/+/g [15,47]. **Ref-Davis17** [16] is a traditional R-VOS benchmark built upon DAVIS17 [32] by providing the language description for a specific object in each video and contains 90 videos with 1,544 expression sentences describing 205 objects in total.

Method	Vision Encoder	Query type	#params	$Decoder^*$	Ref-Davis17			Ref-Youtube-VOS		
					J	F	$J\&F$	J	F	$J\&F$
CMSA [45]	ResNet-50	-	-		32.2	37.2	34.7	33.3	36.5	34.9
CMSA+RNN [45]	ResNet-50	-	-		36.9	43.5	40.2	34.8	38.1	36.4
URVOS [36]	ResNet-50	-	-		47.3	56.0	51.5	45.3	49.2	47.2
ReferFormer [41]	ResNet-50	language	186M		55.8	61.3	58.5	54.8	56.5	55.6
R^2-VOS [18]	ResNet-50	language	186M		57.2	62.4	59.7	56.1	58.4	57.3
ReferFormer [41]	Swin-L	language	360M		57.6	63.4	60.5	60.8	64.0	62.4
SLVP	ViT-B	vision	258M	✓	**57.6**	**64.9**	**61.3** (+0.8)	**62.5**	**66.3**	**64.4**(+2.0)

Table 1. **Comparison** with the state-of-the-art methods on Ref-Davis17 [16] and Ref-Youtube-VOS [36]. $Decoder^*$ represents if the decoder can benefit from self-supervised pre-training.

Figure 3. **Demos** of 'Vision as Query' model on Ref-Davis17 (left) and VISOR (right). Each row has two frames randomly sampled from the same video. The referring expression input to the model is displayed on the top or bottom of each row.

The dataset is split into 60 videos for training and 30 videos for validation. Since there are two annotators and each of them gives the first frame and full-video language description for each referred object, we report the results by averaging the evaluation scores. For a fair comparison, following [41], we also finetune the pre-trained model on RefCOCO/+/g [15, 47] and Ref-Youtube-VOS [36] and then directly test it on Ref-Davis17 without finetuning.

Adopted R-VOS Benchmark EPIC-KITCHENS **VISOR** [6] is a new dataset of pixel annotations and a benchmark suitable for segmenting hands and active objects in egocentric videos. It annotates videos from EPIC-KITCHENS [5] and consists of 174.4K masks from 32.8K frames of 33 kitchens covering 242 entity classes for training and 41.5K masks from 7.7K frames of 24 kitchens covering 182 entity classes for validation. There are 5 unseen kitchens and 9 zero-shot entity classes in the validation. Thus it comes with a new set of challenges not encountered in existing R-VOS benchmarks. It is proposed for the single-modality Semi-VOS task. We adopt it into the R-VOS task by treating the object names as the referring expressions. **Our proposed method serves as the first strong baseline for the R-VOS task on this benchmark.**

Evaluation Metrics. Following the protocol used by [32, 41, 43], we use the following evaluation metrics: region similarity defined by Jaccard Index/Intersection over Union (J), contour accuracy defined by Boundary F-Measure (F) and their average value ($J\&F$).

Reference Performance On the adopted VISOR benchmark, our proposed method serves as the first strong baseline for the R-VOS task. Thus, we also report STM [29] method trained with VISOR and additional COCO [20] data under the relatively easier Semi-VOS task as the reference performance. COCO [20] is used for temporal-based training by synthesizing a video clip of 3 images from random affine transforms.

We also demonstrate when there are fewer pixel-level annotated datasets, our proposed SLVP can still bring non-negligible improvement to the downstream R-VOS task.

4.3. Vision as Query Results

Ref-Davis17 and Ref-Youtube-VOS Benchmarks. The results and demos are in Table. 1, Fig. 3 Fig. 4, and Fig. 5. Even with less number of parameters, our model's performance can surpass Referformer [41] by +0.8 in terms of $J\&F$ on Ref-Davis17, and +2.0 on Ref-Youtube-VOS.

Ablation Study. Table 2 shows the ablation study on description-rich Ref-Davis17. It shows that the self-

Segmentation Training Datasets	Pretrained-Es	Augs	Frozen T-E	*Decoder**	J	F	$J\&F$
Ref-Davis17 [16]					10.4	21.7	16.1
	✓				37.3	41.0	39.2
	✓	✓			40.8	49.2	45.0
	✓	✓	✓		41.7	50.5	46.1
	✓	✓	✓	✓	46.2	55.7	50.5
RefCOCO/g/+ [15,47], Ref-Davis17 [16]	✓	✓	✓	✓	52.8	59.4	56.1
RefCOCO/g/+ [15,47], Ref-Youtube-VOS [36]	✓	✓	✓	✓	**57.6**	**64.9**	**61.3**

Table 2. **Ablation Study** of 'Vision as Query' Model on Ref-Davis17 Benchmark. *Decoder** represents if the decoder uses pre-trained weights from the proposed self-supervised pre-training stage. 'Pretrained-Es' represents pretrained encoders. 'Frozen T-E' represents the frozen text encoder. 'Augs' represents augmentations.

Method	Pretrained-Es	Augs	Frozen T-E	*Decoder**	Segmentation Training Datasets	Task	J	F	$J\&F$
SLVP					VISOR [7]	RVOS	49.7	53.6	51.7
	✓						56.8	60.2	58.5
	✓	✓					66.6	74.7	70.7
	✓	✓	✓				66.4	74.2	70.3
	✓	✓		✓			**70.8**	**78.8**	**74.8**
STM [29] as reference performance					VISOR [7]	VOS	60.6	64.9	62.8
STM [29] as reference performance					MS-COCO [20] + VISOR [7]		**73.6**	**78.0**	**75.8**

Table 3. Performance of 'Vision as Query' Model on VISOR Benchmark.We adopt VISOR benchmark into the more challenging R-VOS task by treating object names as referring expressions. Our proposed 'Vision as Query' method serves as the **first** strong baseline for the R-VOS task. Thus, we also report STM [29] method trained with VISOR and additional COCO [20] data under the relatively easier Semi-VOS task as the reference performance. 'Frozen T-E' represents the frozen text encoder. 'Augs' represents augmentations.

Table 4. **Comparison** of 'Language as Query' and 'Vision as Query' of SLVP architecture on Ref-Davis17 [16]

Method	#params	J	F	$J\&F$
Language as Query	258M	54.1	61.3	57.7
Vision as Query	258M	**57.6**	**64.9**	**61.3**

supervised pre-trained decoder brings non-negligible (+4.4 in terms of $J\&F$) improvement. Besides, the performance of the 'vision as query' model also gains with pre-trained encoders and augmentations. Interestingly, freezing the text encoder during transfer learning brings +1.1 improvement in terms of $J\&F$ on Ref-Davis17. This is because Ref-Davis17 is a relatively small benchmark with longer referring descriptions. Thus finetuning the text encoder may make the model overfit on the training data of Ref-Davis17.

Adopted VISOR Benchmark. The results and demos are in Table. 3 and Fig. 3. Our proposed method serves as the first strong baseline for the R-VOS task on this benchmark. Thus, we also report STM [29] method trained with VISOR and additional COCO [20] data under the relatively easier Semi-VOS task as the reference performance. We can see without the proposed pre-training strategy, our 'vision as query' architecture can already surpass STM [29] trained on VISOR-only which is under the relatively easier Semi-VOS setting. After initializing our decoder with the self-supervised pre-trained weights, the performance is boosted by +4.1 in terms of $J\&F$, which is only 1.0 lower than the performance of STM [29] with COCO [20] as an additional pixel-level training dataset.

Besides, the performance of the 'vision as query' model also gains with pre-trained encoders and augmentations. Interestingly, freezing the text encoder during transfer learning hurts −0.4 in terms of $J\&F$ on the VISOR. This is because VISOR is a relatively large benchmark but with short entity names as referring expressions thus the text encoder won't overfit on the training data during finetuning.

4.4. Language as Query Results

Ref-Davis17 Benchmark. In Table. 4, with the same number of parameters, our 'language as query' model achieves worse performance than our 'vision as query' model, indicating the superiority of the 'vision as query' fusion method under our proposed SLVP architecture.

Adopted VISOR Benchmark. In Table. 5, after initializing our decoder with the self-supervised pre-trained weights, the performance is boosted by +2.0 in terms of $J\&F$. This 'language as query' model also can surpass STM [29] trained on VISOR-only which is under the relatively easier Semi-VOS setting.

Ablation Study. During transfer learning, we further freeze the vision encoder and find performance drops of −4.8 on Ref-Davis17 in Table. 6 and −7.5 in Table. 5 on

Method	Pretrained-Es	Augs	Frozen T-E	Frozen V-E	$Decoder^*$	Segmentation Training Datasets	Task	J	F	$J\&F$
SLVP	✓	✓				VISOR [7]	RVOS	64.3	71.2	67.8
	✓	✓	✓					65.4	73.0	69.2
	✓	✓	✓	✓				59.5	63.8	61.7
	✓	✓	✓		✓			**67.2**	**75.1**	**71.2**
STM [29] as reference performance						VISOR [7]	VOS	60.6	64.9	62.8
STM [29] as reference performance						MS-COCO [20] + VISOR [7]		**73.6**	**78.0**	**75.8**

Table 5. Performance of 'Language as Query' Model on VISOR Benchmark. We adopt the VISOR benchmark into the more challenging R-VOS task by treating object names as referring expressions. Thus, we also report STM [29] method trained with VISOR and additional COCO [20] data under the relatively easier Semi-VOS task as the reference performance. 'Frozen T-E' and 'Frozen V-E' represent the frozen text encoder and vision encoder. 'Augs' represents augmentations.

Segmentation Training Datasets	Pretrained-Es	Augs	Frozen T-E	Frozen V-E	$Decoder^*$	J	F	$J\&F$
Ref-Davis17 [16]	✓					12.5	16.0	14.3
	✓	✓	✓			43.3	51.5	47.4
	✓	✓	✓	✓		39.6	45.5	42.6
	✓	✓	✓		✓	45.2	55.8	50.5
RefCOCO/g/+ [15, 47], Ref-Youtube-VOS [36]	✓	✓	✓		✓	**54.1**	**61.2**	**57.7**

Table 6. Ablation Study of 'Language as Query' Model on Ref-Davis17 Benchmark. $Decoder^*$ represents if the decoder uses pre-trained weights from the proposed self-supervised pre-training stage. 'Pretrained-Es' represents pretrained encoders. 'Frozen T-E' and 'Frozen V-E' represent the frozen text encoder and vision encoder. 'Augs' represents augmentations.

VISOR. This indicates 'language as query' method also relies on strong visual features to perform R-VOS task, indirectly indicating the superiority of the 'vision as query' method. Besides, we still observe $+3.1$ improvement in terms of $J\&F$ brought by the self-supervised pre-trained decoder on Ref-Davis17 in Table. 6 but only $+1.0$ improvement on VISOR in Table. 5. This indicates that the 'language as query' method can benefit from the pre-trained decoder mainly when finetuning on description-rich datasets.

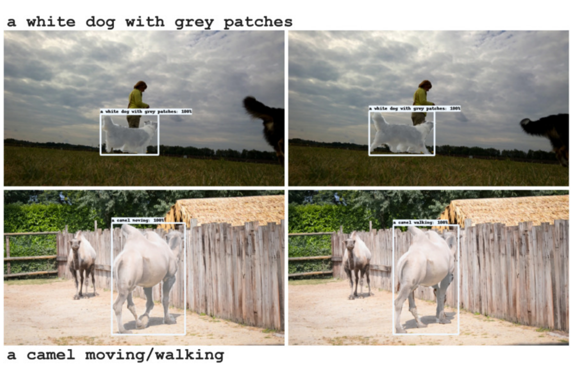


Figure 4. **Demos** of 'Vision as Query' model on four frames of two videos from Ref-Davis17.

Figure 5. **Demos** of 'Vision as Query' model on some random frames from Ref-Davis17.

5. Conclusion

We proposed a general architecture, *i.e.,* SLVP, for the R-VOS task which can support either the 'vision as query' or 'language as query' fusion method. Experiments showed the superiority of the under-studied 'vision as query' method on both description-rich and -poor datasets. Specifically, we presented an effective self-supervised language-vision pre-training strategy to benefit the decoder, enabling non-negligible improvement to the downstream R-VOS task. Existing works do not explore how to make the decoder benefit from the pre-training, leaving it challenging to further improve the performance. Our work is a step in trying to bridge this gap. Besides demonstrating our 'vision as query' model's better performance on well-studied Ref-Davis17 and Ref-Youtube-VOS benchmarks even with fewer model parameters, we further adopt the challenging VISOR benchmark to the R-VOS task. Our 'vision as query' model serves as the first strong baseline. We sincerely acknowledge the inspiring discussions with Chen Sun and Anelia Angelova at Google.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 3

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[4] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems*, 34:11781–11794, 2021. 2

[5] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. 6

[6] 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. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022. 6

[7] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Ely Locke Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 7, 8

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[9] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16321–16330, 2021. 2

[10] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv preprint arXiv:2205.09113*, 2022. 2

[11] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2, 3, 4

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[14] Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. Referring image segmentation via cross-modal progressive comprehension. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10488–10497, 2020. 2

[15] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 5, 6, 7, 8

[16] 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. 5, 6, 7, 8

[17] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 1, 2

[18] Xiang Li, Jinglu Wang, Xiaohao Xu, Xiao Li, Bhiksha Raj, and Yan Lu. Robust referring video object segmentation with cyclic structural consensus. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22236–22245, 2023. 2, 6

[19] Chen Liang, Yu Wu, Tianfei Zhou, Wenguan Wang, Zongxin Yang, Yunchao Wei, and Yi Yang. Rethinking cross-modal interaction from a top-down perspective for referring video object segmentation. *arXiv preprint arXiv:2106.01061*, 2021. 2

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6, 7, 8

[21] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[22] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 10034–10043, 2020. 2

[23] Jie Mei, Jenq-Neng Hwang, Suzanne Romain, Craig Rose, Braden Moore, and Kelsey Magrane. Absolute 3d pose es-

timation and length measurement of severely deformed fish from monocular videos in longline fishing. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2175–2179. IEEE, 2021. 2

[24] Jie Mei, Jenq-Neng Hwang, Suzanne Romain, Craig Rose, Braden Moore, and Kelsey Magrane. Video-based hierarchical species classification for longline fishing monitoring. In *International Conference on Pattern Recognition*, pages 422–433. Springer, 2021. 2

[25] Jie Mei, Suzanne Romain, Craig Rose, Kelsey Magrane, and Jenq-Neng Hwang. Hcil: Hierarchical class incremental learning for longline fishing visual monitoring. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3662–3666. IEEE, 2022. 2

[26] Jie Mei, Jingxi Yu, Suzanne Romain, Craig Rose, Kelsey Magrane, Graeme LeeSon, and Jenq-Neng Hwang. Unsupervised severely deformed mesh reconstruction (dmr) from a single-view image for longline fishing. In *2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2022. 2

[27] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019. 5

[28] Mathew Monfort, SouYoung Jin, Alexander Liu, David Harwath, Rogerio Feris, James Glass, and Aude Oliva. Spoken moments: Learning joint audio-visual representations from video descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14871–14881, June 2021. 4, 5

[29] 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, 6, 7, 8

[30] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[31] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 1

[32] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 5, 6

[33] 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. 1, 2

[34] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 5

[35] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 4

[36] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 208–223. Springer, 2020. 5, 6, 7, 8

[37] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017. 4

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 2

[39] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008. 2

[40] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9481–9490, 2019. 2

[41] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4984, 2022. 2, 3, 5, 6

[42] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 2

[43] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 1, 6

[44] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *Computer Vision–ECCV 2020: 16th European*

Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V, pages 332–348. Springer, 2020. 2, 5

[45] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10502–10511, 2019. 6

[46] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 5

[47] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 5, 6, 7, 8

[48] Haotian Zhang, Yizhou Wang, Zhongyu Jiang, Cheng-Yen Yang, Jie Mei, Jiarui Cai, Jenq-Neng Hwang, Kwang-Ju Kim, and Pyong-Kun Kim. U3d-molts: Unified 3d monocular object localization, tracking and segmentation. In *ICCV Segmenting and Tracking Every Point and Pixel: 6th Workshop on Benchmarking Multi-Target Tracking*, volume 6, 2021. 2

[49] Yuting Zhang, Luyao Yuan, Yijie Guo, Zhiyuan He, I-An Huang, and Honglak Lee. Discriminative bimodal networks for visual localization and detection with natural language queries. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 557–566, 2017. 2

[50] Aotian Zheng, Jie Mei, Farron Wallace, Craig Rose, Rania Hussein, and Jenq-Neng Hwang. Progressive mixup augmented teacher-student learning for unsupervised domain adaptation. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 3030–3034. IEEE, 2023. 2

[51] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3