HTML: Hybrid Temporal-scale Multimodal Learning Framework for Referring Video Object Segmentation

Mingfei Han[1,4], Yali Wang[†2,6], Zhihui Li[3], Lina Yao[4], Xiaojun Chang[1,5], Yu Qiao[6,2]
[1]ReLER, AAII, UTS [2]Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, SIAT-SenseTime Joint Lab, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences [3]Shandong Artificial Intelligence, Qilu University of Technology
[4]Data61, CSIRO [5]Department of Computer Vision, Mohamed bin Zayed University of Artificial Intelligence [6]Shanghai AI Laboratory, Shanghai, China
https://mingfei.info/HTML/

Abstract

Referring Video Object Segmentation (RVOS) is to segment the object instance from a given video, according to the textual description of this object. However, in the open world, the object descriptions are often diversified in contents and flexible in lengths. This leads to the key difficulty in RVOS, i.e., various descriptions of different objects are corresponding to different temporal scales in the video, which is ignored by most existing approaches with single stride of frame sampling. To tackle this problem, we propose a concise Hybrid Temporal-scale Multimodal Learning (HTML) framework, which can effectively align lingual and visual features to discover core object semantics in the video, by learning multimodal interaction hierarchically from different temporal scales. More specifically, we introduce a novel inter-scale multimodal perception module, where the language queries dynamically interact with visual features across temporal scales. It can effectively reduce complex object confusion by passing video context among different scales. Finally, we conduct extensive experiments on the widely used benchmarks, including Ref-Youtube-VOS, Ref-DAVIS17, A2D-Sentences and JHMDB-Sentences, where our HTML achieves state-of-the-art performance on all these datasets.

1. Introduction

Referring Video Object Segmentation (RVOS) has witnessed the growing interest, due to its wide applications in visual editing, virtual reality, human-robotic interaction and so on. Different from the traditional vision-only VOS, RVOS aims to segment the object instance from an input video, according to an open-world description about the referred object. In this case, the model has to learn both visual and textual contents comprehensively, in order to discover the underlying object by multimodal interaction.

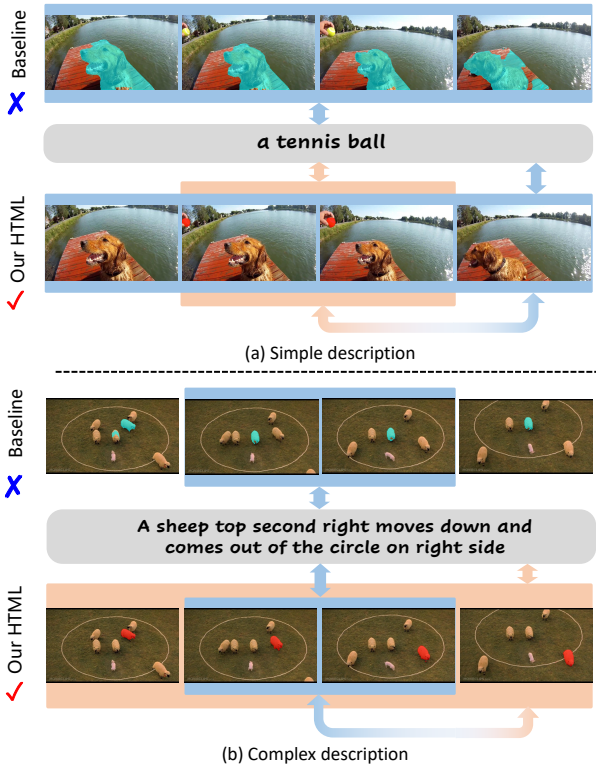

Figure 1: **Referring descriptions in different lengths**. (a) The description is simple containing only the category name. (b) The description is complicated with movement and position of the object. Single-scale baseline (e.g., four frames in (a) and two frames in (b)) fails to segment the referred object, while our hybrid-scale HTML succeeds. More discussion can be found in introduction.

† Corresponding author.

Recent studies [4, 12, 28, 29] have shown that, cross-modal attention is an effective way to bridge the gap between vision and language in RVOS. However, these approaches perform vision-language interactions with video frames sampled from a single temporal scale, which may limit their power to infer the referred object with accurate segmentation. The main reason is that, the open-world descriptions vary in length and contain rich semantics about the referred object, e.g., where it is, how it moves, which objects it interact with. Apparently, such diversified texts are corresponding to various temporal-scale snippets.

For example, the language query in Fig. 1 (a) is *a tennis ball*. Such a short description is corresponding to the ball located at a small region in the middle two frames. If the single-scale baseline samples four frames as input, it will fail to segment the referred object. This is because it overlooks the *dog* in the center place among all these four frames, while lacking the detailed understanding in the middle two frames. Alternatively, the language query in Fig. 1 (b) is *a sheep top second right moves down and comes out of the circle*. Such a long description is corresponding to the particular sheep in the group, which moves across frames. If the single-scale baseline samples two frames as input, it will fail to segment the referred object. This is because it is misled by the subtle movement of sheep group in only two frames, without understanding how each sheep moves from the adjacent frames.

To tackle this difficulty, we propose a concise Hybrid Temporal-scale Multimodal Learning (HTML) Framework for RVOS, which can alleviate object confusion by language-vision interactions across different temporal scales. Specifically, we sample video frames according to different temporal scales. For each temporal scale, we introduce an intra-scale multimodal perception module, which can effectively exploit core visual semantics within the frames at this temporal scale, by mutual enhancement between textual and visual embeddings. Then, we design an inter-scale multimodal perception module, where linguistic embeddings dynamically interact with visual features across temporal scales. In this case, we can hierarchically leverage object context from all the temporal scales to boost RVOS. Finally, we evaluate our HTML on a number of benchmarks, including Ref-Youtube-VOS [22], Ref-DAVIS17 [9], A2D-Sentences and JHMDB-Sentences [7]. The extensive experiments have shown that, our HTML achieves the state-of-the-art performance on all of them.

Overall we make three contributions in this paper:

- Concise and unified learning framework: our Hybrid Temporal-scale Multimodal Learning (HTML) framework hierarchically constructs multimodal interactions via different strides of frame sampling, which can mutually enhance embeddings from both modalities for accurate segmentation.
- Effective multimodal perception module: our Cross-scale Multimodal Perception (CMP) module can effectively reduce complex object confusions with *intra-scale* and *inter-scale* multimodal perceptions, where linguistic and visual features interact across temporal scales.
- State-of-the-art performance on the widely-used benchmarks, which shows the superiority of our framework. Specifically, on Ref-Youtube-VOS [22], our method with ResNet-50 achieves **57.8** in $\mathcal{L}\&\mathcal{F}$, **outperforming the recent SOTA method** [29] **with ResNet-101.**

2. Related works

Vision-only Video Segmentation. Vision-only video segmentation tasks [20, 19, 31, 1, 30, 27, 10, 32, 18, 25, 23, 36], need to segment the objects with predefined semantic query set. Video instance segmentation (VIS) [31] needs to segment the different instances in the desired category set and track each instance with its identity kept. Video object segmentation (VOS) [20, 19], the model needs to separate an object from the background in the video, given the mask of the first mask. In the early stage of VIS, heavy supervision and complex heuristic rules are applied to associate the instances across frames by separate segmentation and association branches, such as MaskTrackRCNN[31] and MaskProp[1]. Recent progress in transformer[24] and DETR[3] has inspired the community with end-to-end works, such as SeqFormer[30] and VisTR[27]. As for VOS, most recent works [32, 18, 25, 23, 36] on semi-suprevised VOS enforces temporal consistency in the video to propagate the first-frame mask to the other frames sequentially. Most of them lie in the group of matching-based methods, updating the memories and matching to segment.

Referring Video Object Segmentation. RVOS task [22] is tasked to segment the referred object from the given video with specified open-world language description. Most early methods in RVOS proposed to refer the object by applying image-level methods on video frames separately and associate them with heuristic rules. However, they usually fail to utilize the temporal dynamic. [22] casts the task as a joint problem of referring segmentation in frame and mask propagation across frames by a memory attention module. [12] proposed a top-down pipeline by constructing exhaustive set of object tracklets and then selecting the target by matching the language features with the all the candidate tracklets. [37] proposed to model the temporal dynamic with an additional optical flow modality. [28] argued the importance of the structural information of video content and proposed to utilize the frame, object and video features simultaneously to obtain better representation. More recently, [2] introduced the DETR structure to RVOS area and [29] proposed to use language-conditional queries to simplify the referring pipeline and improve the performance, which serves as our baseline.

Different from the previous works, we raise the mismatch issue that the various descriptions of different objects are corresponding to different temporal scales of the video. Moreover, we propose a concise HTML framework via multimodal interaction across different temporal scales to capture the core object semantics in the video.

3. Method

To effectively align diversified descriptions and complex videos, we propose a distinct Hybrid Temporal-scale Multimodal Learning (HTML) framework for RVOS. In this section, we introduce our HTML in detail. First, we deliver an overview of HTML framework. Then, we explain how to build the hybrid temporal-scale multimodal learning paths, in the aid of vision-conditioned linguistic decoder and language-conditioned visual decoder. Next, we introduce a Cross-scale Multimodal Perception (CMP) module to align multimodal features across temporal scales. Finally, we describe the training objectives to optimize our HTML.

3.1. Framework Overview

As shown in Fig. 2, our HTML framework consists of three main parts. First, we need to extract visual and linguistic features from backbones. We adopt a visual backbone to extract frame features from T frames sampled from the given video. It can be either 2D CNN networks or 3D transformer networks. We then feed the extracted vision features into a deformable transformer encoder [38] to construct spatiotemporal relations between different frames. Meanwhile, to make a fair comparison with previous works in RVOS [29], we utilize the pretrained linguistic embedding model, RoBERTa [15], to extract textual features $\mathbf{s}_e \in \mathbb{R}^{L\times C}$ from language descriptions with L words. More details can be found in Sec. 4.2.

After extracting visual and linguistic features, we next construct multimodal interactions between the language descriptions and the videos. Different from the previous approaches [29, 28], we build L multimodal learning paths, where the linguistic embedding hierarchically interacts with visual features in different temporal scales. Then, we incorporate the mutually enhanced visual and linguistic features by a novel Cross-scale Multimodal Perception (CMP) module to align multimodal features across different scales. Finally, we design the training losses.

3.2. Hybrid Temporal-scale Multimodal Learning

To capture core object semantics, we propose a novel hybrid temporal-scale multimodal learning framework to learn multimodal relations. To start with, we build hybrid temporal scales via different sampling strides. Then, we construct basic multimodal relation learning units. Finally, we explain how to construct hybrid learning paths.

3.2.1 Hybrid Temporal Scale Construction

Since the texts of various objects may refer to different video parts, single temporal scale often fails to describe the diversified textual contents. To simulate such diversity and flexibility, we build hybrid temporal scales by periodicly sampling frames with different strides.

We first regard all the input frames as the first temporal scale, and then build other $L-1$ temporal scales upon it in a sequential manner. In order to ensure the diversity of sampled temporal scale, we randomly pick one frame from every h frames of last scale, where h denotes the predefined stride. Subsequently, we feed the sampled frames into visual encoder $\mathrm{Encoder(V)}$ to extract feature maps for each of the scales respectively. Specifically, for temporal scale l, we can obtain $\mathbf{M} \in \mathbb{R}^{T\times H\times W\times C}$, where T denotes the number of frames in the temporal scale.

3.2.2 Multimodal Relation Learning

In order to discover the core object semantics, we construct multimodal relations via vision-conditioned linguistic decoder and language-conditioned visual decoder to align semantics between different modalities.

Vision-Conditioned Linguistic Decoder. In order to align linguistic object semantics to the vision contents, we design a vision-conditioned linguistic decoder $\mathrm{Decoder(L|V)}$. Specifically, we have visual features $\mathbf{M} \in \mathbb{R}^{T\times H\times W\times C}$, and linguistic embeddings $\mathbf{s}_e \in \mathbb{R}^{1\times C}$. The vision-conditioned multimodal relations $\mathbf{e} = \mathrm{Decoder}(\mathbf{s}_e|\mathbf{M})$ are constructed by

$$\mathbf{e}_0 = \mathrm{DeformAttn}(\mathbf{s}_e + \mathbf{q}, \mathbf{M}), \quad (1)$$

$$\mathbf{e}_k = \mathrm{DeformAttn}(\mathbf{e}_{k-1}, \mathbf{M}), \quad (2)$$

where $k \in \{1, ..., K-1\}$. We first add $\mathbf{s}_e$ with learnable queries $\mathbf{q} \in \mathbb{R}^{N\times C}$ to represent candidate instances in the video. Then, we use deformable attention module [38] to reason the vision-conditioned multimodal relations where vision features serve as key and value to decompose linguistic features, as in Eq. (1). Finally, we stack the cross-attention module for K times, as in Eq. (2).

Language-Conditioned Visual Decoder. In order to align visual object semantics to linguistic contents, we design a language-conditioned visual decoder $\mathrm{Decoder(V|L)}$ (similar to Eqs. (1) and (2)), to enhance visual representation with the attendance of the language description. Differently, vision features are enhanced by multi-head self-attention (MHSA) modules at the first place, and then linguistic features $\mathbf{s}_e$ serve as key and value in cross-attention modules. In this case, it can reason the language-conditioned multimodal relations. Finally, we can get enhanced visual features by language-conditioned multimodal relations, as $\mathbf{F} = \mathrm{Decoder}(\mathbf{M}|\mathbf{s}_e)$, where $\mathbf{F} \in \mathbb{R}^{T\times H\times W\times C}$.

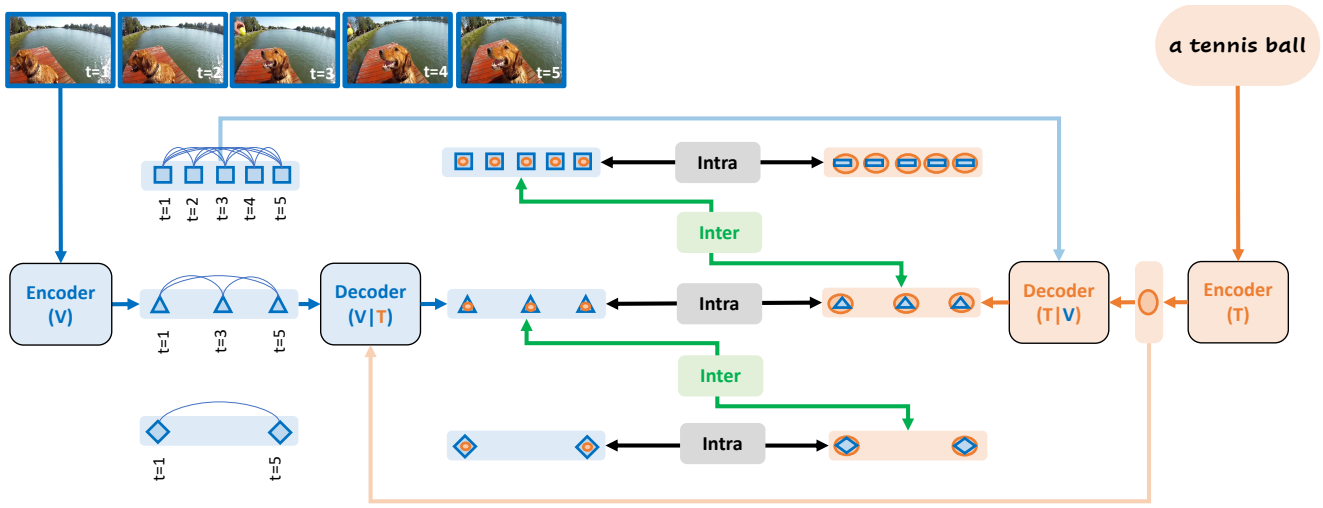

Figure 2: Our Hybrid Temporal-scale Multimodal Learning framework. It aligns linguistic and visual features by learning hierarchical multimodal interactions with hybrid temporal scales, detailed in Sec. 3.2.2. Moreover, a Cross-scale Multimodal Perception (CMP) module is designed to enable interaction and cooperation among temporal scales, detailed in Sec. 3.3.

Hierarchical Multimodal Learning As single-scale multimodal learning is insufficient to understand the relations between videos and texts, we propose to construct the multimodal relations hierarchically for the L hybrid temporal scales with the assistance of $\mathrm{Decoder(L|V)}$ and $\mathrm{Decoder(V|L)}$. We can obtain linguistic-attended visual features and visual-attended linguistic embeddings for different temporal scales, capturing core object semantics conditioned on different visual and linguistic contexts.

3.3. Cross-scale Multimodal Perception

The multimodal relations are constructed conditioned on different modalities with hybrid temporal scales. However, the modeling process of different scales is independent. To promote the cooperation and align the visual and linguistic semantics both within the scale and across the scales, we design Cross-scale Multimodal Perception (CMP) module.

Intra-scale Perception. Despite sharing visual and linguistic feature extraction, the multimodal relation construction via $\mathrm{Decoder(V|L)}$ and $\mathrm{Decoder(L|V)}$ are independent to each other. To promote the cooperation between modalities, we propose an intra-scale perception module.

Specifically, in each temporal scale l, we have visual attended linguistic embeddings $\mathbf{e} \in \mathbb{R}^{N\times T\times D}$ and linguistic attended visual features $\mathbf{F} \in \mathbb{R}^{T\times H\times W\times C}$. To achieve fine-grained semantic alignment, we measure the similarity by dot product between $\mathbf{e}$ and $\mathbf{F}$ on pixel level. Specifically, we obtain the similarity map via multimodal perception module, denoted as $\mathcal{I} = \mathrm{MP}(\mathbf{F}, \mathbf{e})$ by

$$\Omega = \mathrm{MaskHead}(\mathbf{e}), \tag{3}$$

$$\mathcal{I} = \Omega \cdot \mathbf{F}, \tag{4}$$

where $\mathrm{MaskHead}$ denotes three consecutive MLP layers for embedding conversion. Each value of $\mathcal{I}$ represents the relevance between visual-attended linguistic embeddings and linguistic-attended visual features, which can be interpreted as the existence of the referred object. As such, $\mathcal{I}$ is regarded as the object mask prediction with the context of current temporal scale. To this end, we achieve the multimodal perception in the same temporal scale.

Inter-scale Perception. The multimodal relations are constructed in different temporal scales. However, the process in each scale is independent and biased towards the contained object semantics. To alleviate it, we propose to align multimodal features from different temporal scales with a concise inter-scale perception module.

Specifically, suppose that the referred object appears in frame t in temporal scales l and $l+1$, the referred object can be segmented by measuring similarity between $(\mathbf{F}^l(t), \mathbf{e}^l(t))$ and $(\mathbf{F}^{l+1}(t), \mathbf{e}^{l+1}(t))$ simultaneously. Conditioned on same frame t, the visual-attended linguistic embedding $\mathbf{e}^l(t)$ from scale l is supposed to be relevant to the linguistic-attended visual features $\mathbf{F}^{l+1}(t)$ from scale $l+1$. Thus, similar to Eq. (4), the similarity cross different temporal scales can be measured by

$$\mathcal{I}^{l\to l+1}(t) = \mathrm{MP}(\mathbf{e}^l(t), \mathbf{F}^{l+1}(t)). \tag{5}$$

Without losing generality, the inter-scale similarity can also be measured by $\mathcal{I}^{l+1\to l}(t)$. More specifically, frame t can be any frame shared by the adjacent temporal scales, which is ensured by our hybrid temporal scales sampling strategy. Each value in $\mathcal{I}^{l\to l+1}(t)$ represents the referred object prediction with the prior of linguistic object semantic from temporal scale l. In the same way, values in $\mathcal{I}^{l+1\to l}(t)$ rep-

resent the opposite. To this end, we achieve multimodal perception across different temporal scales.

3.4. Training objectives

Our network can be trained in an end-to-end manner to locate and segment the target instance simultaneously. Specifically, the losses for intra-scale and inter-scale multimodal perception in temporal scale l are formed as

$$\mathcal{L}^{l}_{intra} = \sum \mathcal{L}_{cls}(\mathbf{y}, \hat{\mathbf{y}}) + \mathcal{L}_{box}(\mathbf{b}, \hat{\mathbf{b}}) + \mathcal{L}_{mask}(\mathcal{I}, \hat{\mathcal{I}}), \quad (6)$$

$$\mathcal{L}^{l}_{inter} = \sum \mathcal{L}_{mask}(\mathcal{I}^{l+1 \to l}, \hat{\mathcal{I}}) \quad (7)$$

where time and instance subscripts are omitted for simplicity, $\mathbf{y}$ and $\mathbf{b}$ denote binary classification for instance existence and bounding box prediction respectively. Here $\mathcal{L}_{cls}$ is the focal loss [13], $\mathcal{L}_{box}$ is the sum of L1 loss and GIoU loss [21], and $\mathcal{L}_{mask}$ is the combinition of DICE loss [17] and binary mask focal loss. We optimize the network by first finding the best prediction as the positive sample, via minimizing the matching cost $\mathcal{L}^{l}_{intra}$ and $\mathcal{L}^{l}_{inter}$ in each temporal scale l respectively. Then, we average the matching losses from different temporal scales and perception modules, and minimize it for positive samples.

4. Experiments

4.1. Datasets and Metrics

Datasets. We conduct experiments on four datasets: Ref-Youtube-VOS[22], Ref-DAVIS17[9], A2D-Sentences and JHMDB-Sentences[7], following the common practice[29].
Metrics. We follow the standard evaluation protocol [22, 28, 29] to adopt region similarity $\mathcal{L}(\%)$, contour accuracy $\mathcal{F}(\%)$ and mean $\mathcal{L}\&\mathcal{F}$ for Ref-Youtube-VOS and Ref-DAVIS17. For JHMDB-Sentences, we adopt mAP to evaluate the model. For A2D-Sentences, we use Precision@K, Overall IoU, Mean IoU and mAP for evaluation.

4.2. Implemented Details

We set the number of attention layers K to 4 and the hidden dimension C to 256. The number of learnable queries N is set to 5. The number of $\mathrm{MaskHead}$ output channels D is set to 8. During training, we first sample video clips by sliding windows and then generate $L = 3$ hybrid temporal scales with stride $h = 2$ for generalized multimodal representations and relations. We use same training recipes as in [29, 2]. All frames are downsampled by shorter side to 360 and limit the maximumsize for the long side to 640. Our model is pretrained on image referring segmentation datasets [35, 35]. See supplementary materials for details.

During inference, we report results with all input frames in single temporal scale for fair comparisons. On Ref-DAVIS17, we directly inference on models trained on Ref-Youtube-VOS. Similarly, we reports JHMDB-Senteces results directly on models trained with A2D-Sentences.

Method	Backbone	Ref-Youtube-VOS		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
CMSA [33]	ResNet-50	34.9	33.3	36.5
CMSA + RNN [33]	ResNet-50	36.4	34.8	38.1
URVOS [22]	ResNet-50	47.2	45.3	49.2
LBDT-4 [22]	ResNet-50	47.2	45.3	49.2
MLRL [28]	ResNet-50	49.7	48.4	51.0
ReferFormer [29]	ResNet-50	55.6	54.8	56.5
Ours	ResNet-50	**57.8**	**56.5**	**59.0**
PMINet [6]	ResNeSt-101	48.2	46.7	49.6
PMINet + CFBI [6]	ResNeSt-101	53.0	51.5	54.5
CITD [12]	ResNet-101	56.4	54.8	58.1
ReferFormer [29]	ResNet-101	57.3	56.1	58.4
Ours	ResNet-101	**58.5**	**57.3**	**59.8**
PMINet + CFBI [6]	Ensemble	54.2	53.0	55.5
CITD [12]	Ensemble	61.4	60.0	62.7
ReferFormer [29]	Swin-L	62.4	60.8	64.0
Ours	Swin-L	**63.4**	**61.5**	**65.3**
MTTR [2]	Video-Swin-T	55.3	54.0	56.6
ReferFormer [29]	Video-Swin-T	59.4	58.0	60.9
Ours	Video-Swin-T	**61.2**	**59.5**	**63.0**
ReferFormer [29]	Video-Swin-S	60.1	58.6	61.6
Ours	Video-Swin-S	**61.4**	**59.9**	**62.9**
ReferFormer [29]	Video-Swin-B	62.9	61.3	64.6
Ours	Video-Swin-B	**63.4**	**61.5**	**65.2**

Table 1: Comparison with the SOTA methods on Ref-YTB-VOS.

Method	Backbone	Ref-DAVIS17		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
CMSA [33]	ResNet-50	34.7	32.2	37.2
CMSA + RNN [33]	ResNet-50	40.2	36.9	43.5
URVOS [22]	ResNet-50	51.5	47.3	56.0
LBDT-4 [5]	ResNet-50	54.5	-	-
MLRL [28]	ResNet-50	58.0	53.9	62.0
ReferFormer [29]	ResNet-50	58.5	55.8	61.3
Ours	ResNet-50	**59.5**	**56.6**	**62.4**
ReferFormer [29]	Swin-L	60.5	57.6	63.4
Ours	Swin-L	**61.6**	**58.9**	**64.4**
ReferFormer [29]	Video-Swin-B	61.1	58.1	64.1
Ours	Video-Swin-B	**62.1**	**59.2**	**65.1**

Table 2: Comparison with the SOTA methods on Ref-DAVIS17.

4.3. SOTA Comparisons

We compare our method with the state-of-the-art methods on Ref-Youtube-VOS, Ref-DAVIS17, A2D-Sentences and JHMDB-Sentences. On Ref-Youtube-VOS, our approach achieves 58.5 in $\mathcal{L}\&\mathcal{F}(\%)$ with ResNet-50, as shown in Tab. 1, which surpasses the recent SOTA method ReferFormer[29] with same backbone by 2.2 points. Moreover, it surpasses all the other SOTA methods with larger ResNet-101 on all evaluation metrics, which fully sug-

Method	Backbone	Precision					IoU		mAP
		P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	Overall	Mean	
Hu *et al*. [8]	VGG-16	34.8	23.6	13.3	3.3	0.1	47.4	35.0	13.2
Gavrilyuk *et al*. [7]	I3D	47.5	34.7	21.1	8.0	0.2	53.6	42.1	19.8
CMSA + CFSA [34]	ResNet-101	48.7	43.1	35.8	23.1	5.2	61.8	43.2	-
ACAN [26]	I3D	55.7	45.9	31.9	16.0	2.0	60.1	49.0	27.4
CMPC-V [14]	I3D	65.5	59.2	50.6	34.2	9.8	65.3	57.3	40.4
ClawCraneNet [11]	ResNet-50/101	70.4	67.7	61.7	48.9	17.1	63.1	59.9	-
MTTR ($\omega = 8$) [2]	Video-Swin-T	72.1	68.4	60.7	45.6	16.4	70.2	61.8	44.7
MTTR ($\omega = 10$) [2]	Video-Swin-T	75.4	71.2	63.8	48.5	16.9	72.0	64.0	46.1
ReferFormer [29]	Video-Swin-T	82.8	79.2	72.3	55.3	19.3	77.6	69.6	52.8
ReferFormer [29]	Video-Swin-B	83.1	80.4	74.1	57.9	21.2	78.6	70.3	55.0
Ours	Video-Swin-T	82.2	79.2	72.3	55.3	20.1	77.6	69.2	53.4
Ours	Video-Swin-B	**84.0**	**81.5**	**75.8**	**59.2**	**22.8**	**79.5**	**71.2**	**56.7**

Table 3: Comparison with the state-of-the-art methods on A2D-Sentences.

Method	Backbone	mAP
Hu *et al*. [8]	VGG-16	17.8
Gavrilyuk *et al*. [7]	I3D	23.3
ACAN [26]	I3D	28.9
CMPC-V [14]	I3D	34.2
MTTR ($\omega = 8$) [2]	Video-Swin-T	36.6
MTTR ($\omega = 10$) [2]	Video-Swin-T	39.2
ReferFormer† ($\omega = 6$) [29]	Video-Swin-T	39.1
ReferFormer [29]	Video-Swin-T	42.2
ReferFormer [29]	Video-Swin-B	43.7
Ours	Video-Swin-T	42.7
Ours	Video-Swin-B	**44.2**

Table 4: SOTA results comparison on JHMDB-Sentences.

	Components	#Frames	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
i.	Baseline	5	55.6	54.8	56.5
ii.	HTML w/o CMP	5	56.0	54.7	57.4
iii.	HTML	5	56.3	55.0	57.5
iv.	Baseline	8	56.2	55.0	57.3
v.	HTML w/o CMP	8	57.1	56.0	58.2
vi.	HTML	8	**57.8**	**56.5**	**59.0**

Table 5: Ablation study on the components of our HTML.

gests the superiority of our method. When equipped with larger backbone, our method still show considerable superiority with accuracy gap of 1.2 points for ResNet-101 and 1.0 points for Swin-L. We also experiment our method with the well-known Video Swin Transformers [16]. Our method with Video-Swin-Tiny backbone surpasses the SOTA method with the same backbone by 1.8 points. With larger Video-Swin Transformers (Small and Base models), our method still achieves SOTA performance, which shows the generality of our method.

As shown in Tab. 2, our method surpasses the SOTA methods on Ref-DAVIS17 by over 1.0 points on both ResNet-50, Swin-L and Video-Swin-Base backbones, with new a SOTA record 62.1 in $\mathcal{J}\&\mathcal{F}(\%)$. We also experiment our method on A2D-Sentences and JHMDB-Sentences datasets and compare with other SOTA results as shown in Tab. 3 and Tab. 4. Our method achieves SOTA performances with new records on both the two datasets. On A2D-Sentences, our HTML surpass SOTA result by 1.7 points in mAP and higher recall by 1.6 on Precision@0.9. On JHMDB-Sentences, our method still achieves a new SOTA record with 44.2 in mAP. These results demonstrate the superiority of our method.

4.4. Ablation Study

In this section, we ablate core components of our HTML with Ref-Youtube-VOS based on ResNet-50.

Effectiveness of our HTML. To validate the effectiveness of our Hybrid Temporal-scale Multimodal Learning framework, we investigate each of our components by gradually adding them to the baseline [29]. First, comparing (i)&(iii) in Tab. 5, our HTML improves the performance of baseline to 56.3 when 5 frames are used for training, which is better than the baseline trained with longer input frames. Further, when longer temporal input is available, comparing (iv)&(vi), our HTML improves the model by 1.6 to 57.8 in $\mathcal{J}\&\mathcal{F}$. These prove the effectiveness of our method with frames in different lengths.

Second, comparing counterpart networks using different number of frames, *i.e*. (iii) and (vi), our method can benefit from longer temporal input (8 frames vs. 5 frames) with an improvement of 1.5 points. This conclusion holds among the other counterpart network pair, *i.e*. (ii)&(v). Further, each of our components brings an improvement of 0.3 when trained with total 5 frames, while the improvement can be doubled to 0.7-0.9 with total 8 frames. This indicates that our method can better utilize the long temporal input.

Finally, taking 8 frames for instance, hierarchical multimodal learning and cross-scale multimodal perception improves $\mathcal{J}\&\mathcal{F}$ by 0.9 and 0.7 respectively. It proves the ef-

#Scales	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
1	56.2	55.0	57.3
2	56.9	55.8	58.1
3	**57.8**	**56.5**	**59.0**

(a) Effect of No. of temporal scales.

Frames	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
5	56.3	55.0	57.5
8	**57.8**	**56.5**	**59.0**
12	57.5	56.4	58.6

(b) Effect of No. of input frames.

Inter-scale	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
None	56.2	55.0	57.3
$l \rightarrow l+1$	56.9	55.8	58.0
$l+1 \rightarrow l$	**57.8**	**56.5**	**59.0**

(c) Effect of direction of CMP.

Table 6: Ablation study on Hybrid Temporal Scales in (a) and (b), and Cross-scale Multimodal Perception in (c).

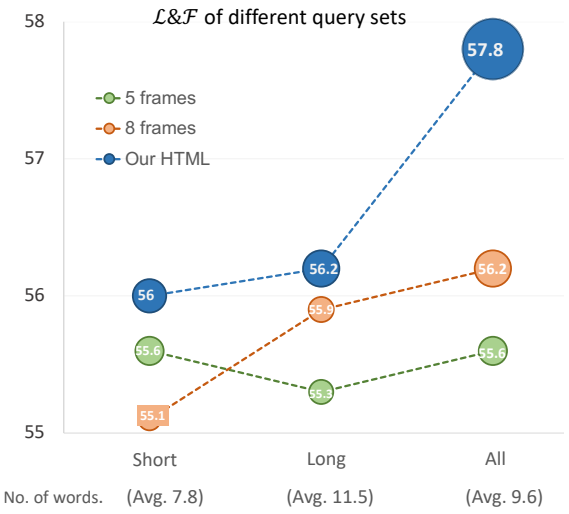

Figure 3: Performance comparison of different query sets with different temporal scales on Ref-Youtbe-VOS.

fectiveness of each component of our HTML.

Length of language descriptions. We explore the performance of our method with different sampled sets of language descriptions, to validate the ability of capturing diversified linguistic object semantics. Since the train set of Ref-Youtube-VOS contains an average of two language descriptions for each object, we sample the shorter ones to form the Short set and the others to form the Long set.

As shown in bottom line of Fig. 3, the Long set has longer sentences than the Short set in the average number of words. As shown in Fig. 3, when the objects are described by Short query set, more input frames in single temporal scale achieve inferior performance (5 frames vs. 8 frames: 55.6 vs. 55.1). It's interesting to note that more visual content (8 frames) fails to improve the performance when the linguistic content is insufficient (shorter queries). When the complexity of queries increases, *i.e.* with the Long query set, the quantitative relation is reversed: 8 frames guided model obtains better performance than the model with 5 frames, increasing by 0.8 (8 frames-Short query vs. 8 frames-Long query: 55.1 vs. 55.9). We infer that this is caused by the mismatch of visual and linguistic object semantics. On one side, shorter quires, *i.e.* relatively simple linguistic semantics, are insufficient to interpret longer videos. On the other side, the longer language descriptions contain more content irrelevant to the visual input.

Differently, our models trained with either the Short set or the Long set all surpass the single temporal scale guided model. Impressively, when the objects are described by queries flexible in lengths, *i.e.* with the All set, our method gets a performance boost by 1.6, while single temporal scale baseline (both 8 frames and 5 frames guided networks) improves by 0.3. This shows that our method has the strong ability to solve the mismatch issue between visual content and diversified linguistic contents.

Number of hybrid temporal scales. We investigate the effectiveness of our hierarchical multimodal learning, by exploring the number of hybrid temporal scales. For fair comparison among different settings, the number of total input frames is set to 8 in this subsection. As shown in Tab. 6 (a), when two temporal scales are constructed, our method brings an improvement of 0.6 in $\mathcal{L}\&\mathcal{F}$. When the number of temporal scales is increased to three, continuous improvement of 0.9 (56.9 vs. 57.8) is observed. It proves that our model benefits from the increasing visual diversity constructed by hybrid temporal scales.

Number of input frames. We explore the effect of total input frames here. In this subsection, hybrid temporal scales are constructed by default following Sec. 3.2.2. As shown in Tab. 6 (b), more input frames bring an improvement by 1.5 in $\mathcal{L}\&\mathcal{F}$ (5 frames vs. 8frames: 56.3 vs. 57.8). When the input frames increase continuously to 12 frames, the performance saturates and drops slightly to 57.5. We conjecture that it is caused by insufficient video-language pairs (8 frames vs. 12 frames: 49k vs. 32k) compared to largely increased computation complexity (8 frames vs. 12 frames: 21.5 GFLOPs vs. 33.6 GFLOPs for transformers).

Direction of inter-scale perception. We explore the effect of direction of inter-scale multimodal perception in CMP. Comparing first two lines of Tab. 6 (c), $l \rightarrow l+1$ perception improves the baseline by 0.7 points; As in first and last lines, $l+1 \rightarrow l$ perception improves the baseline by 1.6 points. These prove the effectiveness of our inter-scale perception. We choose the latter one for better performance.

4.5. Visualizations

We visualize the segmentation results of complex and simple language descriptions in Fig. 4. Specifically, we

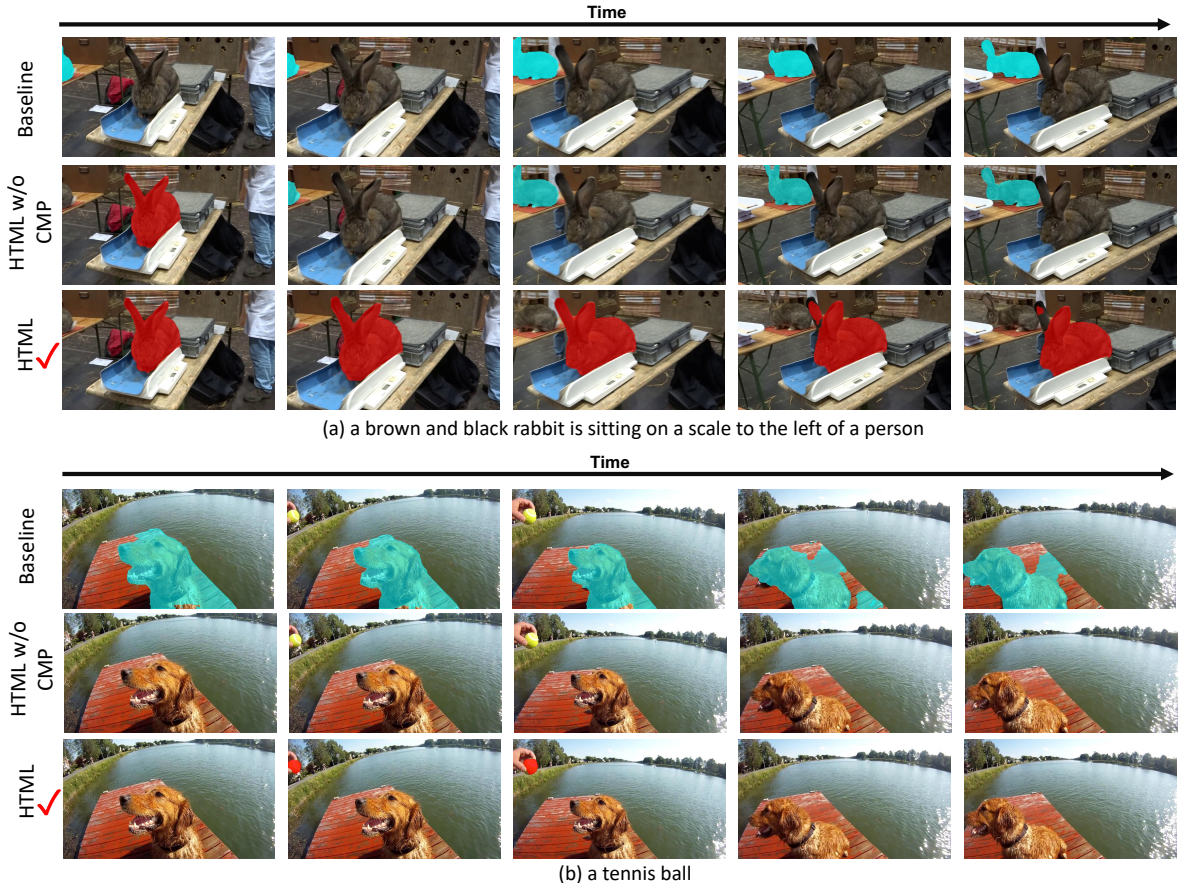

Figure 4: Visualization results of complex and simple language descriptions on Ref-Youtube-VOS. Red masks indicate positive segmentation results and blue masks indicate the negatives. Our HTML is able to clarify such object confusion.

compare three settings, *i.e.*, baseline with only single temporal scale, HTML w/o CMP which constructs hybrid temporal scales, and our final HTML which dynamically construct multimodal relations cross temporal scales. As expected, when only single temporal scale is adopted, baseline fails to segment the target in the video, *e.g.*, the background rabbit in the subplot (a) is mistakenly referred. The main reason is that the rabbit shares similar appearance to the target object and also locates *to the left of a person* in last three frames. The model is misled and confused by the single temporal scale. When applied with hybrid temporal scales, the language description can interact with both long and short temporal scales. Thus, the previous false prediction in the first frame is corrected. Further, when applied with CMP, our model is able to clarify the object confusion by discovering the core semantics *on a scale* and make correct predictions. Similarly, in subplot (b), baseline is misled by the video clip where tennis ball only appears in the middle two frames. When gradually applying our proposed modules, the mistakenly predicted dog is clarified and further the target tennis ball is correctly segmented.

5. Conclusion

In this work, we develop a HTML framework to align linguistic and visual features by learning multimodal relations hierarchically in different temporal scales. Moreover, we introduce an inter-scale multimodal perception module to construct dynamic multimodal interactions across temporal scales. We conduct experiments on four datasets and establish new state-of-the-art results. Particularly, our method with ResNet-50 backbone surpasses the recent methods with ResNet-100. The comprehensive ablation experiments and visualization results show that our method is able to discover core object semantics in the different modalities.

Acknowledgement

This work was supported in part by the National Key R&D Program of China (No.2022ZD0160100, No. 2022ZD0160505), the National Natural Science Foundation of China under Grant (62272450), the Joint Lab of CAS-HK, the Shenzhen Research Program (RCJC20200 714114557087), and the Youth Innovation Promotion Association of Chinese Academy of Sciences (No.2020355)

References

[1] Gedas Bertasius and Lorenzo Torresani. Classifying, segmenting, and tracking object instances in video with mask propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9739–9748, 2020. 2

[2] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. *arXiv preprint arXiv:2111.14821*, 2021. 2, 5, 6

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 2

[4] 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

[5] Zihan Ding, Tianrui Hui, Junshi Huang, Xiaoming Wei, Jizhong Han, and Si Liu. Language-bridged spatial-temporal interaction for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4964–4973, 2022. 5

[6] Zihan Ding, Tianrui Hui, Shaofei Huang, Si Liu, Xuan Luo, Junshi Huang, and Xiaoming Wei. Progressive multimodal interaction network for referring video object segmentation. *The 3rd Large-scale Video Object Segmentation Challenge*, page 7, 2021. 5

[7] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and action video segmentation from a sentence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5958–5966, 2018. 2, 5, 6

[8] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *European Conference on Computer Vision*, pages 108–124. Springer, 2016. 6

[9] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Asian Conference on Computer Vision*, pages 123–141. Springer, 2018. 2, 5

[10] Liulei Li, Wenguan Wang, Tianfei Zhou, Jianwu Li, and Yi Yang. Unified mask embedding and correspondence learning for self-supervised video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18706–18716, 2023. 2

[11] Chen Liang, Yu Wu, Yawei Luo, and Yi Yang. Clawcranenet: Leveraging object-level relation for text-based video segmentation. *arXiv preprint arXiv:2103.10702*, 2021. 6

[12] 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, 5

[13] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 5

[14] Si Liu, Tianrui Hui, Shaofei Huang, Yunchao Wei, Bo Li, and Guanbin Li. Cross-modal progressive comprehension for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 6

[15] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 3

[16] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021. 6

[17] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. 5

[18] 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

[19] 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. 2

[20] 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. 2

[21] 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. 5

[22] 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. 2, 5

[23] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 2

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017. 2

[25] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmenta-

tion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9481–9490, 2019. 2

[26] Hao Wang, Cheng Deng, Junchi Yan, and Dacheng Tao. Asymmetric cross-guided attention network for actor and action video segmentation from natural language query. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3939–3948, 2019. 6

[27] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8741–8750, 2021. 2

[28] Dongming Wu, Xingping Dong, Ling Shao, and Jianbing Shen. Multi-level representation learning with semantic alignment for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4996–5005, 2022. 2, 3, 5

[29] 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

[30] Junfeng Wu, Yi Jiang, Wenqing Zhang, Xiang Bai, and Song Bai. Seqformer: a frustratingly simple model for video instance segmentation. *arXiv preprint arXiv:2112.08275*, 2021. 2

[31] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5188–5197, 2019. 2

[32] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *European Conference on Computer Vision*, pages 332–348. Springer, 2020. 2

[33] 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. 5

[34] Linwei Ye, Mrigank Rochan, Zhi Liu, Xiaoqin Zhang, and Yang Wang. Referring segmentation in images and videos with cross-modal self-attention network. *arXiv preprint arXiv:2102.04762*, 2021. 6

[35] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016. 5

[36] Yurong Zhang, Liulei Li, Wenguan Wang, Rong Xie, Li Song, and Wenjun Zhang. Boosting video object segmentation via space-time correspondence learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2246–2256, 2023. 2

[37] Wangbo Zhao, Kai Wang, Xiangxiang Chu, Fuzhao Xue, Xinchao Wang, and Yang You. Modeling motion with multi-modal features for text-based video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11737–11746, 2022. 2

[38] 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