

Robust Referring Video Object Segmentation with Cyclic Structural Consensus

Xiang Li^{1*}, Jinglu Wang², Xiaohao Xu³, Xiao Li², Bhiksha Raj^{1,4}, Yan Lu²
¹Carnegie Mellon University, ²Microsoft Research Asia, ³University of Michigan,
⁴Mohamed bin Zayed University of Artificial Intelligence

Abstract

Referring Video Object Segmentation (R-VOS) is a challenging task that aims to segment an object in a video based on a linguistic expression. Most existing R-VOS methods have a critical assumption: the object referred to must appear in the video. This assumption, which we refer to as “**semantic consensus**”, is often violated in real-world scenarios, where the expression may be queried against false videos. In this work, we highlight the need for a robust R-VOS model that can handle semantic mismatches. Accordingly, we propose an extended task called Robust R-VOS (R^2 -VOS), which accepts unpaired video-text inputs. We tackle this problem by jointly modeling the primary R-VOS problem and its dual (text reconstruction). A **structural text-to-text cycle constraint** is introduced to discriminate semantic consensus between video-text pairs and impose it in positive pairs, thereby achieving multi-modal alignment from both positive and negative pairs. Our structural constraint effectively addresses the challenge posed by linguistic diversity, overcoming the limitations of previous methods that relied on the point-wise constraint. A new evaluation dataset, R^2 -Youtube-VOS is constructed to measure the model robustness. Our model achieves state-of-the-art performance on R-VOS benchmarks, Ref-DAVIS17 and Ref-Youtube-VOS, and also our R^2 -Youtube-VOS dataset.

1. Introduction

Referring video object segmentation (R-VOS) aims to segment a referred object in a video given a linguistic expression. R-VOS has witnessed growing interest thanks to its promising potential in human-computer interaction, such as video editing and augmented reality.

In previous studies [1, 40, 7, 3], the R-VOS problem is addressed with the strict assumption that the referred object appears in the video, thus requiring object-level **semantic consensus** between the expression and the video. However, this assumption is not always true in practice, leading

*This work was done when Xiang Li and Xiaohao Xu were interns at MSRA.

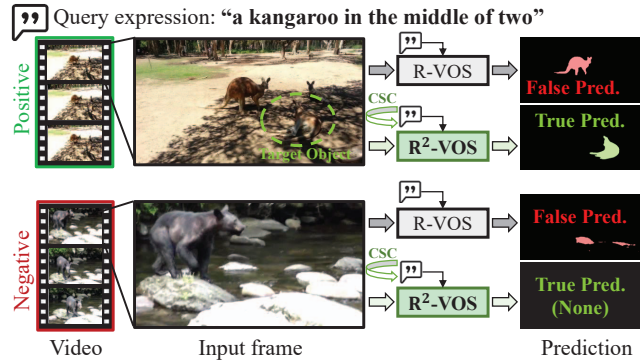


Figure 1. **Illustration of the new R^2 -VOS task.** A linguistic expression is given to query a set of videos without the semantic consensus assumption in R^2 -VOS. Videos containing the referred object by the expression are **positive**, otherwise **negative**. Existing R-VOS methods such as Referformer [40] easily segment irrelevant objects in both positive and negative videos, as they do not investigate the consensus. In contrast, our R^2 -VOS method achieves greater accuracy by incorporating cyclic structural consensus (CSC).

to severe false-alarm problems in negative videos (referred object does not present in the video), as demonstrated in Fig. 1. This limitation hinders the application of such methods in scenarios where matched vision-language pairs are not available. We argue that the current R-VOS task is incomplete when assuming that the referred object always exists in the video.

Even with semantic consensus in positive video-language pairs, locating the correct object is still challenging due to its multimodal nature. Recent method MTTR [1] employs multimodal transformer encoders to learn a joint representation, and localizes the object by ranking all presented objects in the video. ReferFormer [40] and the image method ReTR [15] adopt the linguistic expression as a query to the transformer decoder to avoid redundant ranking of all objects. However, these latest R-VOS methods still suffer from a semantic misalignment of the segmented object and the linguistic expression, as shown in Fig. 1. Some previous works [30, 2, 5] attempt to enhance the alignment by imposing cycle consistency between referring expressions and their reconstructed counterparts, with-

out considering negative vision-language pairs. Moreover, they restrict the two expressions to be identical or similar, which may not always be the case due to linguistic diversity. Expressions that describe an object can vary from different perspectives, e.g., the target kangaroo in Fig. 1 can be described either as “in the middle of two” or “lying on its side”. Therefore, a more robust cycle constraint is required to better accommodate this diversity.

In this paper, we seek to investigate the semantic alignment problem between visual and linguistic modalities in referring video segmentation. We extend the current R-VOS task to accept unpaired video-language inputs. This new task, termed **Robust R-VOS (R²-VOS)**, goes beyond the limitation of the existing R-VOS task by analyzing the semantic consensus. It additionally discriminates between **positive** video-language pairs where the referred object appears in the video, and **negative** pairs where it does not.

The R²-VOS task essentially corresponds to two interrelated problems: the **primary** problem of segmenting masks in videos with referring texts (R-VOS), and its **dual** problem of reconstructing text expressions from videos with object masks. To jointly optimize both problems, consistency in the text-to-text cycle is usually imposed if the video-language pairs are positive, as in most previous works [30, 2, 5]. We go further by learning from both positive and negative pairs. Specifically, we investigate the semantic consensus using cycle consistency measure to discriminate the positive and negative pairs, while also imposing consistency in positive pairs. Both the discrimination and imposing procedure help with the video-language alignment learning. Furthermore, instead of using point-wise cycle consistency for individual samples [2, 5], we adopt a **structural** measurement to assess the relation consistency between the referring textual space and its reconstructed counterpart. This design accommodates linguistic diversity better, as it considers that an object may have different text expressions but should be differentiated from the expressions describing other objects. In addition, we enable the end-to-end joint training of the primary and dual problems by introducing an object localizing module (OLM) as a proxy to bridge the two problems. The dual text reconstruction can be conducted using the proxy features without waiting for the resulting masks from the primary task.

Our contributions are summarized as:

- We are the first to address the severe false-alarm problem faced by previous R-VOS methods with unpaired video-language inputs. We introduce the new R²-VOS task accepting unpaired inputs, as well as an evaluation dataset and corresponding metrics.
- We introduce the cyclic structural consensus for both discrimination between positive and negative video-language pairs and enhancement of segmentation quality, which better accommodates linguistic diversity.

- We propose a R²-VOS network that enables end-to-end training of the primary referring segmentation and dual expression reconstruction task by introducing the object localizing module as a proxy.
- Our method achieves state-of-the-art performance for both R-VOS and R²-VOS tasks on popular Ref-Youtube-VOS, Ref-DAVIS, and new R²-Youtube-VOS datasets.

2. Related Works

Referring segmentation. Referring image segmentation (RIS) and R-VOS aim to segment an object in an image or video sequence respectively, given a linguistic description as the query. Recent RIS methods [42, 6] achieves promising results by using the power of multimodal transformer. Going beyond the RIS task [10, 16, 44, 23, 29, 41, 28], R-VOS is a more challenging task as it requires leveraging both intra-frame and temporal cues. URVOS [34] is the first unified R-VOS framework with a cross-modal attention and a memory attention module, which largely improves R-VOS performance. ClawCraneNet [18] leverages cross-modal attention to bridge the semantic correlation between textual and visual modalities. ReferFormer [40] and MTTR [1] are two latest works that utilize transformers to decode or fuse multimodal features. ReferFormer [40] employs a linguistic prior to the transformer decoder to focus on the referred object. MTTR [1] leverages a multimodal transformer encoder to fuse linguistic and visual features. Different from other vision-language tasks, e.g., image-text retrieval [22, 24, 31] and video question answering [13, 36], R-VOS needs to construct object-level multimodal semantic consensus in a dense visual representation.

Consistency constraint in multi-modal learning. The pioneer work [30] jointly model generation and comprehension of linguistic object description in images and find it benefits both tasks. KACNet [2] improves visual grounding by enhancing semantic alignment with knowledge-aided consistency constraints. [5] also utilize caption-aware consistency to improve referring image segmentation. All these methods use text-to-text (point-wise) consistency, assuming that the referring text and reconstructed text are similar. However, in the real world, an object can be described with diverse expressions from different perspectives. In our work, we do not enforce the cycle consistency between two expressions but rather the structural consistency between two expression spaces. Moreover, all these methods only utilize cycle consistency to enhance aligning multi-modal features. They tend to fail when the referred object is not presented in the image, as discussed in [30].

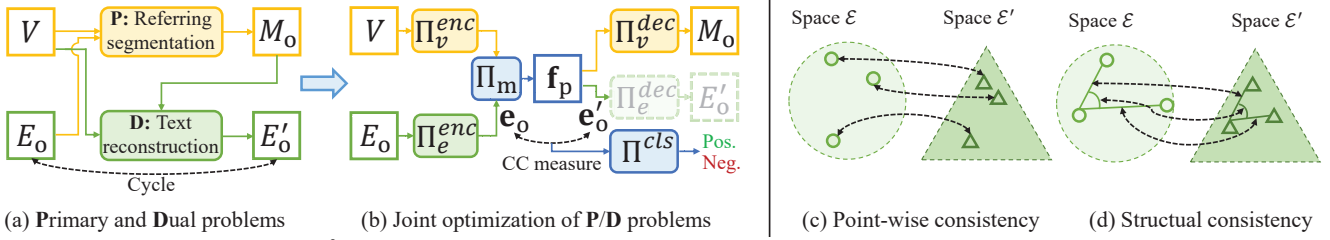


Figure 2. **Problem analysis.** (a) R^2 -VOS explores the **Primary** problem of referring segmentation and the **Dual** problem of text reconstruction. The **P/D** problems are linked in a cycle path from the input expression E_o to the reconstructed expression E'_o . (b) We utilize the cycle consistency (CC) measure for 1) discriminating positive/negative pairs with module Π^{cls} , 2) alignment enhancement for positive pairs. We enable the joint optimization of **P/D** problems with a cross-modal proxy f_p between all uni-modal operations (i.e., Π_v^{enc} , Π_e^{enc} , Π_v^{dec} and Π_e^{dec}). (c) Point-wise consistency is not suitable in R^2 -VOS as the mapping between \mathcal{E} and \mathcal{E}' are not necessarily bijective (an object can be described with various expressions). (d) Instead, we apply a structural consistency to preserve relationships (distances and angles).

3. Robust R-VOS

3.1. Problem Definition

We introduce a novel task, robust referring video segmentation (R^2 -VOS), which aims to predict mask sequences $\{M_o\}$ for an unconstrained video set $\{V\}$ given an expression E_o of an object o . Different from the setting of the previous R-VOS problem, R^2 -VOS does not impose the constraint that the queried videos must contain the object specified by the expression E_o . A video V and an expression E_o have **semantic consensus** if the object o referred to by E_o appears in V . Then, the video is considered *positive* with respect to E_o , otherwise, it is *negative*. In the R^2 -VOS task, the goal is to discriminate between positive and negative videos, and predict masks M_o of the target object o for positive videos. For negative videos, all frames are treated as background.

Primary problem. The referring segmentation can be formulated as the maximum *a posteriori* estimation problem of $p(M_o|V, E_o)$. By applying the Bayes rule, we obtain:

$$p(M_o|V, E_o) \sim p(E_o|V, M_o)p(M_o|V) \quad (1)$$

Dual problem. As the prior $p(M_o|V)$ is not affected by the expression E_o , we consider maximizing $p(E_o|V, M_o)$ as a dual problem of the referring segmentation (primary problem), which is to reconstruct the text expression given the video and object masks.

Joint optimization. The primary and dual problem can be linked in a cyclic path, as previous works [35, 5], between the input expression E_o and the reconstructed expression E'_o , as depicted in Fig. 2 (a). Previous methods only consider positive text-video pairs and impose consistency between E_o and E'_o for all samples without differentiating between positive and negative pairs. To go a step further, we consider that the negative pairs can also contribute to joint optimization. We model the semantic consensus for each text-video pair with a consistency measure that allows the text-video alignment to benefit from imposing cycle consistency in positive pairs and also from discrimination between

positive and negative pairs.

To avoid the impact of pretrained language models and enable end-to-end training, we investigate the consistency measurement between the original textual embedding space \mathcal{E} and the reconstructed textual embedding space \mathcal{E}' . To link the primary and dual problem and facilitate the joint optimization, we introduce a cross-modal proxy feature f_p that captures information from both the inputs of the primary and dual problems, i.e., (V, E_o) and (V, M_o) , after multi-modal interaction Π_m , as shown in Fig. 2 (b). f_p is placed between the unimodal encoders and decoders, i.e., Π_v^{enc} , Π_e^{enc} , Π_v^{dec} , Π_e^{dec} , to focus solely on the multimodal interaction. An additional module Π^{cls} for discriminating between positive and negative pairs.

3.2. Cyclic Consensus Measurement

As mentioned, we utilize the text-to-text cycle consistency to measure the text-video semantic consensus.

Point-wise cycle consistency. Previous works compute point-wise cycle consistency for each individual sample (Fig. 2 (c)), i.e., $CC(e_i, e'_i), \forall e_i \in \mathcal{E}, e'_i \in \mathcal{E}'$, where CC denotes a distance measure function. However, we observe that the mapping between the two spaces \mathcal{E} and \mathcal{E}' is not necessarily bijective, as there could be multiple textual descriptions for the same object from different perspectives. As a result, e_i and e'_i may not be close. Thus, simply utilizing point-wise consistency will collapse the feature space to be sub-optimal.

Structural cycle consistency. In contrast, we consider that the structure of the reconstructed embedding space \mathcal{E}' should be preserved as that of \mathcal{E} . Although an object may have different text expressions, they should be differentiated from the expressions describing other objects. Thus, the relationship between expressions that describe different objects should be preserved, as illustrated in Fig. 2 (d). Mathematically, the structure-preserving property is defined as isometric and conformal constraints to preserve pair-wise

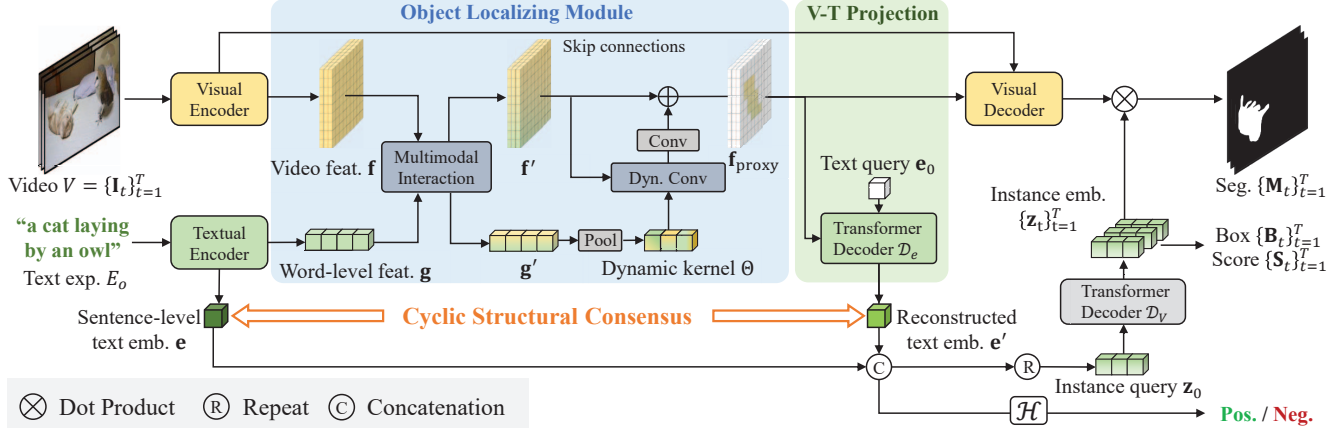


Figure 3. **Overview of the R²-VOS network.** Given a video clip $V = \{\mathbf{I}_t\}_{t=1}^T$ and a textual expression E_o referring to object o , we first extract unimodal features \mathbf{f} and \mathbf{g} , then fuse them in the object localizing module to obtain the proxy feature $\mathbf{f}_{\text{proxy}}$. Then, $\mathbf{f}_{\text{proxy}}$ is projected to a textual space as \mathbf{e}' . The semantic consensus assessed with cyclic structural consistency between \mathbf{e}' and \mathbf{e} is used for discriminating positive/negative pairs with \mathcal{H} , and imposed for positive pairs. The final segmentation is obtained by dynamic convolutions with features from the visual decoder and dynamic weights from \mathbf{e}' and \mathbf{e} .

distances and angles for $\mathbf{e} \in \mathcal{E}$ and $\mathbf{e}' \in \mathcal{E}'$:

$$CC_{dist}(|\mathbf{e}_1 - \mathbf{e}_2|, |\mathbf{e}'_1 - \mathbf{e}'_2|), \quad (2)$$

$$CC_{angle}(\angle(\overrightarrow{\mathbf{e}_2\mathbf{e}_1}, \overrightarrow{\mathbf{e}_2\mathbf{e}_3}), \angle(\overrightarrow{\mathbf{e}'_2\mathbf{e}'_1}, \overrightarrow{\mathbf{e}'_2\mathbf{e}'_3})), \quad (3)$$

where $|\cdot|$ denotes a distance metric, CC_{dist} and CC_{angle} denote cycle consistency measurement function for distances and angles, respectively. The detailed implementation is elaborated in the loss function (Section 4.5).

4. R²-VOS Network

In this section, we elaborate on the structure of R²-VOS network, which mainly consists of four parts: feature extraction, object localizing module (OLM), video-text (V-T) projection for text reconstruction, and mask decoding for final segmentation, as illustrated in Fig. 3. We first extract the video feature \mathbf{f} , word-level text feature \mathbf{g} , and sentence-level text feature \mathbf{e} . On the one hand, to model the primary segmentation problem of maximizing $p(M_o|V, E_o)$, we enable the multimodal interaction in OLM to generate the grounded feature $\mathbf{f}_{\text{proxy}}$. $\mathbf{f}_{\text{proxy}}$ coarsely locates the referred object o with irrelevant features suppressed, which serves as a proxy linking the primary segmentation and dual text reconstruction problem. The final mask M_o is obtained by dynamic convolution [4] on the decoded visual feature maps, with kernels learned from instance embedding $\{\mathbf{z}_t\}_{t=1}^T$. On the other hand, to model the dual text reconstruction problem of maximizing $p(E_o|V, M_o)$, we utilize $\mathbf{f}_{\text{proxy}}$ as the alternative of V and M_o , since $\mathbf{f}_{\text{proxy}}$ conveys grounded video clues of object o . In this way, we enable the parallel optimization of the primary and dual problem by relating them to $\mathbf{f}_{\text{proxy}}$. Specifically, we employ a V-T projection module to project $\mathbf{f}_{\text{proxy}}$ onto a

reconstructed text embedding \mathbf{e}' . We add a cyclic structural constraint between \mathbf{e}' and \mathbf{e} to enforce the semantic alignment between the segmented mask and expression for positive videos. In addition, we introduce a classification head $\mathcal{H}(\mathbf{e}, \mathbf{e}')$, which discriminates the semantic consensus by assessing the consistency between \mathbf{e} and \mathbf{e}' .

4.1. Unimodal Feature Extraction

Visual encoder. We build the visual encoder with a visual backbone and a deformable transformer encoder [46] on top of it, as that in [40]. The extracted features from the backbone are flattened, projected to a lower dimension, added with positional encoding [11], and then fed into a deformable transformer encoder [46] similar to the previous method [40]. We denote the multi-scale output of the transformer encoder as \mathbf{F} and the low-resolution visual feature map from the backbone as \mathbf{f} , where $\mathbf{f} \in \mathbb{R}^{T \times C_v \times \frac{H}{32} \times \frac{W}{32}}$, C_v is the feature channel, T is the video length and H and W are the original image size.

Textual encoder. We leverage a pre-trained linguistic model RoBERTa [25] to map the input textual expression E_o to a textual embedding space. The textual encoder extracts a sequence of word-level text feature $\mathbf{g} \in \mathbb{R}^{C_e \times L}$ and a sentence-level text embedding $\mathbf{e} \in \mathbb{R}^{C_e \times 1}$, where C_e and L are the dimension of linguistic embedding space and the expression length respectively.

4.2. Object Localizing Module

We employ the object localizing module (OLM) to coarsely locate the referred object o . The grounded feature $\mathbf{f}_{\text{proxy}}$ encoding information of o can not only be utilized for the primary segmentation problem, but also for the dual expression reconstruction task, which serves as a proxy con-

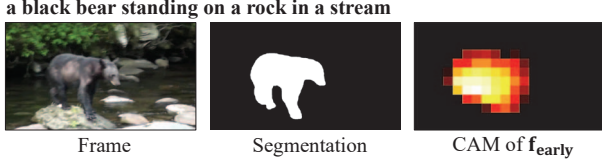


Figure 4. Channel activation map (CAM) of $\mathbf{f}_{\text{proxy}}$.

necting the two problems. Specifically, we utilize dynamic convolution [4] to discriminate visual features in the early stage. As shown in the blue part of Fig. 3, we first enable the multimodal interaction between video and text features, then apply the dynamic convolution with kernels learned from text features to discriminate the object-level semantics. In particular, multi-head cross-attention (MCA) [37] is leveraged to conduct the multimodal interaction:

$$\begin{aligned} \mathbf{h}_f &= \text{LN}(\text{MCA}(\mathbf{f}, \mathbf{g}) + \mathbf{f}) & \mathbf{f}' &= \text{LN}(\text{FFN}(\mathbf{h}_f) + \mathbf{h}_f) \\ \mathbf{h}_g &= \text{LN}(\text{MCA}(\mathbf{g}, \mathbf{f}) + \mathbf{g}) & \mathbf{g}' &= \text{LN}(\text{FFN}(\mathbf{h}_g) + \mathbf{h}_g) \end{aligned} \quad (4)$$

where $\text{MCA}(\mathbf{X}, \mathbf{Y})$ denotes $\text{Attention}(\mathbf{W}^Q\mathbf{X}, \mathbf{W}^K\mathbf{Y}, \mathbf{W}^V\mathbf{Y})$. \mathbf{W} represents learnable weight. LN and FFN denote layer normalization and feed-forward network respectively. The word-level text feature \mathbf{g}' is further pooled to a fixed length, and followed by a fully-connected layer to form the word kernels $\Theta = \{\theta_i\}_{i=1}^K$. K is the kernel number and $\theta_i \in \mathbb{R}^{C \times 1}$. The dynamic kernels are applied separately to video feature $\mathbf{f}' \in \mathbb{R}^{C \times THW}$ to form the $\mathbf{f}_{\text{proxy}} \in \mathbb{R}^{C \times THW}$:

$$\mathbf{f}_{\text{proxy}} = \text{BN}(\varphi(\theta_1 \mathbf{f}' \oplus \dots \oplus \theta_K \mathbf{f}') + \mathbf{f}'), \quad (5)$$

where \oplus is the concatenation in channel dimension and $\varphi(\cdot)$ is a 1×1 convolution to reduce the dimension. BN denotes batch normalization. Fig. 4 shows a visualization of $\mathbf{f}_{\text{proxy}}$. The channel activation map of $\mathbf{f}_{\text{proxy}}$ indicates that the OLM can effectively locate the referred object and suppress irrelevant features.

4.3. Text Reconstruction

V-T projection. We leverage a transformer decoder \mathcal{D}_E as the textual decoder to transform the visual representation of the referred object into the textual space. As shown in Fig. 3, a learnable text query $\mathbf{e}_0 \in \mathbb{R}^{C_e \times 1}$ is employed to query the $\mathbf{f}_{\text{proxy}}$. The output of the transformer decoder is the reconstructed text embedding $\mathbf{e}' = \mathcal{D}_E(\mathbf{f}_{\text{proxy}}, \mathbf{e}_0) \in \mathbb{R}^{C_e \times 1}$.

4.4. Referring Segmentation

Mask segmentation. We leverage deformable transformer decoders with dynamic convolution to segment the object masks as that in [40]. Since the reconstructed text embedding is generated with visual features injected, we consider it can encode some visual information, thus augmenting the original text embedding. As shown in Fig. 3, we

first fuse the reconstructed text embedding \mathbf{e}' to text embedding \mathbf{e} . The fused text embedding \mathbf{e} is then repeated N times to form the instance query [38] $\mathbf{z}_0 \in \mathbb{R}^{C_q \times N}$, where C_q is the dimension of instance query and N is the instance query number. We then use $T \times$ deformable transformer decoders \mathcal{D}_V with shared weights to decode the instance embeddings $\mathbf{z}_t \in \mathbb{R}^{C_q \times N}$ for each frame, i.e., $\mathbf{z}_t = \mathcal{D}_V(\mathbf{F}_t, \mathbf{z}_0)$. \mathbf{F}_t is the multiscale visual feature from visual encoder at time t . A dynamic kernel \mathbf{w}_t is further learned from the instance embedding \mathbf{z}_t . The final feature map $\mathbf{f}_{\text{out},t} \in \mathbb{R}^{C \times H \times W}$ is obtained by fusing low-level features from the feature pyramid network [20] in the visual decoder. The mask prediction $\mathbf{M}_t \in \mathbb{R}^{N \times H \times W}$ can be computed by $\mathbf{M}_t = \mathbf{w}_t^T \mathbf{f}_{\text{out},t}$.

Auxiliary heads. We build a set of auxiliary heads to obtain the final object masks across frames. In particular, a box head, a scoring head and a semantic consensus discrimination head are leveraged to predict the bounding boxes $\mathbf{B}_t \in \mathbb{R}^{N \times 4}$, confidence scores $\mathbf{S}_t \in \mathbb{R}^{N \times 1}$ and the alignment degree of multimodal semantics $A \in \mathbb{R}$. The box and scoring head are two fully-connected layers upon the instance embedding \mathbf{e}_t . The semantic consensus discrimination head $\mathcal{H}(\mathbf{e}, \mathbf{e}')$ consists of two fully-connected layers upon the text embeddings $\mathbf{e} \oplus \mathbf{e}'$. Note that A represents the semantic alignment in the entire video rather a single frame, since the expression is a video-level description.

4.5. Loss Function

The loss function of our method can be boiled down to three parts:

$$\mathcal{L} = \lambda_{\text{text}} \mathcal{L}_{\text{text}} + \lambda_{\text{segm}} \mathcal{L}_{\text{segm}} + \lambda_{\text{align}} \mathcal{L}_{\text{align}}, \quad (6)$$

where $\mathcal{L}_{\text{text}}$, $\mathcal{L}_{\text{segm}}$, and $\mathcal{L}_{\text{align}}$ are losses for text reconstruction, referring segmentation and semantic consensus discrimination respectively.

Loss of consensus discrimination. $\mathcal{L}_{\text{align}}$ is a cross-entropy loss between the predicted alignment A and ground-truth \hat{A} . $\hat{A} = 1$ represents positive video-language pairs, otherwise 0.

Loss for text reconstruction. Given the text embedding \mathbf{e} and reconstructed embedding \mathbf{e}' , we use a structural constraint to impose the cycle consistency between \mathbf{e} and \mathbf{e}' for positive pairs. We calculate the loss by

$$\mathcal{L}_{\text{text}} = \hat{A} \cdot (\mathcal{L}_{\text{dist}} + \lambda_{\text{angle}} \mathcal{L}_{\text{angle}}), \quad (7)$$

where λ_{angle} is a hyperparameter balancing the distance loss $\mathcal{L}_{\text{dist}}$ and angle loss $\mathcal{L}_{\text{angle}}$. We elaborate these two losses according to the structural cycle consistency n in Section 3.2. Let $\mathcal{X}^n = \{(x_1, \dots, x_n) | x_i \in \mathcal{X}\}$ denote a set of n -tuples, $\Phi^n = \{(x, x') | x \in \mathcal{X}^n, x' \in \mathcal{X}'^n\}$

denote a set of pairs consisting of two n -tuples of distinct elements from two different sets \mathcal{X} and \mathcal{X}' . Specifically, the distance-based and angle-based relations relate text embeddings of 2-tuple and 3-tuple respectively, i.e., $\Phi^2 = \{(\mathbf{x}, \mathbf{x}') | \mathbf{x} = (\mathbf{e}_i, \mathbf{e}_j), \mathbf{x}' = (\mathbf{e}'_i, \mathbf{e}'_j), i \neq j\}$ and $\Phi^3 = \{(\mathbf{x}, \mathbf{x}') | \mathbf{x} = (\mathbf{e}_i, \mathbf{e}_j, \mathbf{e}_k), \mathbf{x}' = (\mathbf{e}'_i, \mathbf{e}'_j, \mathbf{e}'_k), i \neq j \neq k\}$. Then the losses are given by:

$$\mathcal{L}_{dist} = \sum_{(\mathbf{x}, \mathbf{x}') \in \Phi^2} l_\delta(\phi_D(\mathbf{x}'), \phi_D(\mathbf{x})) \quad (8)$$

$$\mathcal{L}_{angle} = \sum_{(\mathbf{x}, \mathbf{x}') \in \Phi^3} l_\delta(\phi_\angle(\mathbf{x}), \phi_\angle(\mathbf{x}')) \quad (9)$$

where $\phi_D(\mathbf{x}) = \frac{1}{\mu(\mathbf{x})} \|\mathbf{e}_i - \mathbf{e}_j\|_2$, $\phi_\angle(\mathbf{x}) = \cos\angle(\mathbf{e}_i, \mathbf{e}_j, \mathbf{e}_k)$ and $\mu(\mathbf{x}) = \sum_{\mathbf{x}=(x_1, x_2) \in \mathcal{X}^2} \frac{\|x_1 - x_2\|_2}{|\mathcal{X}^2|}$ is the average distance function. The Huber loss $l_\delta(x, x') = \frac{1}{2}(x - x')^2$ if $|x - x'| \leq 1$, otherwise $|x - x'| - \frac{1}{2}$.

Loss for referring segmentation. Given a set of predictions $\mathbf{y} = \{\mathbf{y}_i\}_{i=1}^N$ and ground-truth $\hat{\mathbf{y}}$, where $\mathbf{y}_i = \{\mathbf{B}_{i,t}, \mathbf{S}_{i,t}, \mathbf{M}_{i,t}\}_{t=1}^T$ and $\hat{\mathbf{y}} = \{\hat{\mathbf{B}}_t, \hat{\mathbf{S}}_t, \hat{\mathbf{M}}_t\}_{t=1}^T$, we search for an assignment $\sigma \in \mathcal{P}_N$ with the highest similarity where \mathcal{P}_N is a set of permutations of N elements ($\hat{\mathbf{y}}$ is padded with 0). The similarity can be computed as

$$\mathcal{L}_{match}(\mathbf{y}_i, \hat{\mathbf{y}}) = \lambda_{box} \mathcal{L}_{box} + \lambda_{conf} \mathcal{L}_{conf} + \lambda_{mask} \mathcal{L}_{mask}, \quad (10)$$

where λ_{box} , λ_{conf} , and λ_{mask} are weights to balance losses. Following previous works [8, 38], we leverage a combination of Dice [17] and BCE loss as \mathcal{L}_{mask} , focal loss [21] as \mathcal{L}_{conf} , and GIoU [33] and L1 loss as \mathcal{L}_{box} . The best assignment $\hat{\sigma}$ is solved by the Hungarian algorithm [12]. Given the best assignment $\hat{\sigma}$, the segmentation loss between ground-truth and predictions is defined as $\mathcal{L}_{segm} = \mathbb{1}(\hat{A}) \cdot \mathcal{L}_{match}(\mathbf{y}, \hat{\mathbf{y}}_{\hat{\sigma}(i)})$.

4.6. Inference

During inference, we select the candidate with the highest confidence to predict the final masks:

$$\{\bar{\mathbf{M}}_t\}_{t=1}^T = \{\mathbb{1}(A > 0.5) \cdot \mathbf{M}_{\bar{s}, t}\}_{t=1}^T \quad (11)$$

$$\bar{s} = \arg \max_i \{\mathbf{S}_{i,1} + \dots + \mathbf{S}_{i,T}\}_{i=1}^N,$$

where $\{\bar{\mathbf{M}}_t\}_{t=1}^T$ is the masks of referred object. $\mathbf{S}_{i,t}$ and $\mathbf{M}_{i,t}$ represent the i -th candidate in \mathbf{S}_t and \mathbf{M}_t . \bar{s} is the slot with the highest confidence to be the target object. We use $\mathbb{1}(A)$ to filter out predictions in negative videos to mitigate false alarms. $\mathbb{1}(A > 0.5) = 1$ if $A > 0.5$, else 0.

5. Experiment

5.1. Dataset and Metrics

Dataset. We conduct experiments on three datasets: Ref-Youtube-VOS [34], Ref-DAVIS [32] and R²-Youtube-VOS.

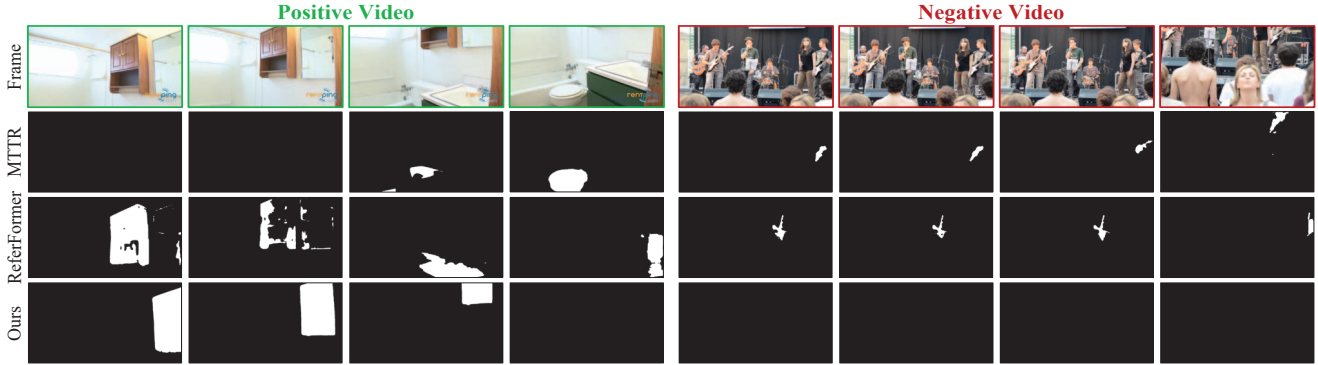
We construct a new **evaluation** dataset, R²-Youtube-VOS, which extends the Ref-Youtube-VOS validation set with each object containing both positive and negative linguistic queries. We generate diverse negative text-video pairs by either replacing object properties {category, action, appearance, position} in the original positive expression or randomly picking another irrelevant expression to the object. In this way, we construct the negative expressions with different granularity (each video has about 10 negative expressions). The positive text-video pairs are the same as in Ref-Youtube-VOS and we evaluate the segmentation quality on such positive text-video pairs.

Metrics. We inherit the commonly-used region similarity \mathcal{J} and contour accuracy \mathcal{F} [32] metrics for segmentation quality evaluation in R-VOS and R²-VOS tasks. Note that R-VOS evaluate \mathcal{J} and \mathcal{F} scores with pure positive pairs, which could be different from R²-VOS, because we need to additionally predict if the pairs are positive first in R²-VOS. We introduce two additional metrics for robustness evaluation in R²-VOS tasks:

- **Semantic alignment accuracy \mathcal{A} .** \mathcal{A} is the binary classification accuracy for semantic alignment.
- **Robustness score \mathcal{R} .** The alignment accuracy may not comprehensively represent the robustness quality, since even with false alignment, we can still get the right masks using mask confidence (For example, a negative pair is misclassified as positive, but the segmentation masks are predicted with very low confidence, and we still get the correct empty masks). Therefore, we propose the new metric $\mathcal{R} = 1 - \frac{\sum_{M \in \mathcal{M}_{neg}} |M|}{\sum_{M \in \mathcal{M}_{pos}} |M|}$ to evaluate the degree of misclassified pixels in negative videos, where \mathcal{M}_{neg} and \mathcal{M}_{pos} are the sets containing segmented masks in negative and positive videos respectively. $|M|$ denotes the foreground area of mask M . The total foreground area of positive videos $\sum_{M \in \mathcal{M}_{pos}} |M|$ serves as a normalization term. Ideally, a model should treat all the negative videos as backgrounds with $\mathcal{R} = 1$.

5.2. Implementation Details

Following previous methods [8, 40], our model is first pre-trained on Ref-COCO+/g dataset [45, 30] and then finetuned on Ref-Youtube-VOS. The model is trained for 6 epochs with a learning rate multiplier of 0.1 at the 3rd and the 5th epoch. The initial learning rate is 1e-4 and a learning rate multiplier of 0.5 is applied to the backbone. We adopt a batchsize of 8 and an AdamW [27] optimizer with weight decay 1×10^{-4} . Following convention [1], the evaluation on Ref-DAVIS directly uses models trained on Ref-Youtube-VOS without re-training. All images are cropped to have the longest side of 640 pixels and the shortest side of 360 pixels during evaluation. The window size is set to



Expression: the mirror in the bathroom is to the right of the wood cabinet

Figure 5. **Qualitative comparison** to the state-of-the-art R-VOS method on the R^2 -VOS task.

Method	P	Backbone	Ref-Youtube-VOS			Ref-DAVIS-17		
			$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}	$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}
Spatial Visual Backbone								
CMSA [43]		ResNet50	34.9	33.3	36.5	34.7	32.2	37.2
CMSA + RNN [43]		ResNet50	36.4	34.8	38.1	40.2	36.9	43.5
URVOS [34]		ResNet50	47.2	45.3	49.2	51.5	47.3	56.0
LBDT [7]		LBDT	49.4	48.2	50.6	54.5	-	-
Wu <i>et al.</i> [39]		ResNet50	49.7	48.4	51.0	57.9	53.9	62.0
PMINet [8]		ResNet101	53.0	51.5	54.5	-	-	-
YOFO [14]		ResNet50	54.4	50.1	58.7	52.2	47.5	56.8
CITD [19]		ResNet101	56.4	54.8	58.1	-	-	-
ReferFormer [40]	✓	ResNet50	55.6	54.8	56.5	58.5	55.8	61.3
R²-VOS	✓	ResNet50	57.3	56.1	58.4	59.7	57.2	62.4
ReferFormer [40]	✓	Swin-T	58.7	57.6	59.9	-	-	-
R²-VOS	✓	Swin-T	60.2	58.9	61.5	-	-	-
Spatio-temporal Visual Backbone								
MTTR [1]		V-Swin-T	55.3	54.0	56.6	-	-	-
Chen <i>et al.</i> [3]		V-Swin-T	55.6	54.0	55.6	-	-	-
ReferFormer [40]		V-Swin-T	56.0	54.8	57.3	-	-	-
R²-VOS		V-Swin-T	57.1	55.9	58.2	-	-	-
ReferFormer [40]	✓	V-Swin-T	59.4	58.0	60.9	-	-	-
R²-VOS	✓	V-Swin-T	61.3	59.6	63.1	-	-	-

Table 1. **Quantitative comparison to state-of-the-art R-VOS methods.** “P” denotes models pretrained on ref-coco+ dataset.

5 for all backbones. We train model with negative samples created by randomly mismatching vision-language pairs in Ref-Youtube-VOS. We set the $\lambda_{text} = 0.1$, $\lambda_{cls} = 2$, $\lambda_{mask} = 2$, $\lambda_{align} = 1$, $\lambda_{angle} = 10$, $\lambda_{L1} = 5$, $\lambda_{giou} = 2$, $\lambda_{dice} = 2$ and $\lambda_{focal} = 5$ during all training process.

5.3. Main Results

Comparison on R-VOS. In Table 1, we first compare our method on Ref-Youtube-VOS. For results of ResNet [9] backbone, our method achieves 57.3 $\mathcal{J}\&\mathcal{F}$ which outperforms the latest method ReferFormer [40] by 1.7 $\mathcal{J}\&\mathcal{F}$. In addition, our method runs at 30 FPS compared to 22 FPS of state-of-the-art ReferFormer (FPS is measured using a single NVIDIA P40 with $batchsize = 1$). For results of Swin-Transformer [26] backbones, our method achieves 60.2 $\mathcal{J}\&\mathcal{F}$ and 61.3 $\mathcal{J}\&\mathcal{F}$ with Swin-Tiny and Video-Swin-Tiny backbones respectively, which outperforms ReferFormer [40] and MTTR [1] by a clear margin. Our method achieves 59.7 $\mathcal{J}\&\mathcal{F}$ on Ref-DAVIS-17 dataset, which outperforms ReferFormer by 1.2 $\mathcal{J}\&\mathcal{F}$. The per-

Method	Segm. Quality			Robust. Quality	
	$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}	\mathcal{R}	\mathcal{A}
ReferFormer [40]	42.2	41.2	43.2	46.3	54.4
MTTR [1]	50.0	50.5	51.5	41.4	49.3
R²-VOS	57.2	56.0	58.4	51.2	64.5

Table 2. **Quantitative comparison to state-of-the-art R-VOS methods for R^2 -VOS task.**

formance improvement on R-VOS task indicates that the semantic consensus consolidation from cyclic training can benefit the segmentation result.

Comparison on R^2 -VOS. As shown in Table 2, we compare our method with state-of-the-art method MTTR [1] and ReferFormer [40] on R^2 -VOS task. We extend the scoring head of MTTR and ReferFormer to discriminate semantic alignment and retrain the model with the same dataset setting as ours. We evaluate the R -VOS task in terms of both segmentation quality and robustness quality. Generally, the segmentation quality of all methods drops compared to the result of R-VOS task. This can be accounted for two reasons. First, the positive videos can be misclassified into negative videos thus the whole video will be segmented as background. Thereby, the semantic discrimination capability will also influence the segmentation quality. Second, for MTTR and ReferFormer, the network design only focuses on positive video segmentation. For example, MTTR requires segmenting all objects in the video and then ranking them. However, by feeding unpaired inputs, the network training can be distracted. Compared to the large drop of $\mathcal{J}\&\mathcal{F}$ by MTTR and ReferFormer, the segmentation quality of our method is slightly influenced. For the robustness quality, our method achieves a 51.2 \mathcal{R} and 64.5 \mathcal{A} which outperforms MTTR and ReferFormer by a clear margin.

Qualitative results. We compare the qualitative results of our method against state-of-the-art methods in Fig. 5 on R^2 -VOS. For **positive videos**: The result indicates that our method predicts accurate and temporally consistent results, while ReferFormer [40] and MTTR [1] fail to locate the correct object. For **negative videos**: Both ReferFormer and

Components	$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}	\mathcal{R}	\mathcal{A}
R-VOS Task					
Baseline	52.4	51.9	52.8	-	-
+OLM	55.5	54.4	56.5	-	-
+OLM+CC	56.9	55.7	58.1	-	-
+OLM+CC+FT	57.3	56.1	58.4	-	-
R ² -VOS Task					
Baseline	41.4	40.8	42.0	48.5	53.2
+OLM+CC	56.8	55.7	58.0	50.0	62.1
+OLM+CC+FT	57.2	56.0	58.4	51.2	64.5

Table 3. **Impact of different components in our method.** OLM: Object localizing module, CC: Consistency constraint, FT: Fusing text embeddings.

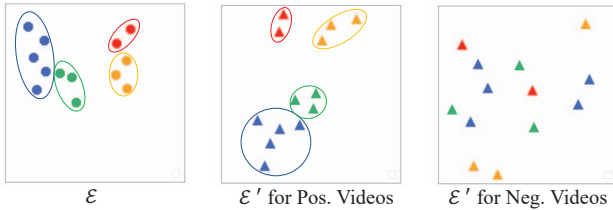


Figure 6. **Visualization of text embedding spaces.** Dots represent original text embeddings in \mathcal{E} , and triangles represent reconstructed ones in \mathcal{E}' induced by positive (the middle sub-figure) and negative (the right sub-figure) videos. Elements in the same color belong to the same object. Note that an object can have multiple text descriptions. The structure of \mathcal{E}' is well preserved from \mathcal{E} for positive videos (ellipses bound embeddings of same objects), while it is not preserved for negative videos.

MTTR suffer from a severe false-alarm problem when the referred object does not exist in the video. In contrast, with multi-modal cycle constraint and consensus discrimination, our method successfully filters out negative videos and mitigates the false alarm. To further explore how consensus discrimination works, we visualize the text embedding and reconstructed text embedding spaces for both positive and negative videos. As shown in Fig. 6, we notice that, for embeddings of positive videos, they preserve relative relations well, while for negative videos, the reconstructed embeddings have a random pattern in the space.

5.4. Ablation Study

Module effectiveness. We conduct ablation study with the ResNet-50 backbone on Ref-Youtube-VOS and R²-Youtube-VOS to investigate the contribution of each component. We build a transformer-based baseline model and equip our proposed components step-by-step. As shown in Table 3, we first evaluate the baseline model on R-VOS task which achieves a 52.4 $\mathcal{J}\&\mathcal{F}$. After employing the OLM, the performance boosts to 55.5 $\mathcal{J}\&\mathcal{F}$ and the full model achieves the best performance of 57.3 $\mathcal{J}\&\mathcal{F}$. For R²-VOS task, since the baseline does not have a cycle consensus criterion, it shows low robustness. By equipping the consistency constraint and text embedding fusing, our method achieves the best (57.2 $\mathcal{J}\&\mathcal{F}$, 51.2 \mathcal{R} and 64.5 \mathcal{A}).

CC	$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}	\mathcal{R}	\mathcal{A}
PW	52.4	51.2	53.6	47.2	51.1
SA	56.3	55.1	57.5	49.8	60.0
SD	56.0	55.0	57.0	47.2	59.3
SD+SA	56.8	55.7	58.0	50.0	62.1

Table 4. **Impact of the cycle consistency (CC) constraint.** PW: Point-wise. SA: Structural angle. SD: Structural distance.

Query Number	$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}	\mathcal{R}	\mathcal{A}
1	54.9	54.2	55.6	51.9	63.4
5	57.2	56.0	58.4	51.2	64.5
9	57.0	56.8	57.2	50.5	64.1

Table 5. **Impact of the query number.**

Consistency constraint. We conduct experiments to ablate the influence of cycle consistency constraints. As shown in Table 4, utilizing point-wise consistency constraint shows an inferior performance compared to the setting with structural constraint. We consider the point-wise constraint may force an injective mapping from the textual domain to the visual domain. However, the mapping can be a many-to-one function for R-VOS, i.e., each object corresponds to multiple textual descriptions. In addition, since the OLM leverages the text feature to locate the referred object, if we use the direct point-wise constraint to form reconstructed text embedding, it will encourage the network to memorize the text feature in the $\mathbf{f}_{\text{proxy}}$ and result in a collapse for text reconstruction. Table 4 shows that the model achieves the best result of 56.8 $\mathcal{J}\&\mathcal{F}$, 50.0 \mathcal{R} and 62.1 \mathcal{A} when with both structural angle and structural distance constraint. Note that we disable the text embedding fusion in this experiment.

Instance query number. Though only one object is referred to for each frame for both R-VOS and R²-VOS tasks, to help the network optimization, we employ more than one instance query to each video. Table 5 indicates that a query number of 5 brings the best overall segmentation quality.

6. Conclusion

In this paper, we investigate the semantic misalignment problem in R-VOS task. A pipeline jointly models the referring segmentation and text reconstruction problem, equipped with a structural cycle consistency constraint, is introduced to discriminate and enhance the semantic consensus between visual and textual modalities. For model robustness evaluation, we extend the R-VOS task to accept unpaired inputs and collect a corresponding R²-Youtube-VOS dataset. We observe a severe false-alarm problem suffered from previous methods on R²-Youtube-VOS while ours accurately discriminates unpaired inputs and segments high-quality masks for paired inputs. Our method achieves state-of-the-art performance on R-VOS benchmarks and R²-Youtube-VOS dataset. We believe that, with unpaired inputs, R²-VOS is a more general setting, which can shed light on a new direction to investigate the robustness of referring segmentation.

References

- [1] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multi-modal transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4985–4995, 2022. 1, 2, 6, 7
- [2] Kan Chen, Jiyang Gao, and Ram Nevatia. Knowledge aided consistency for weakly supervised phrase grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4042–4050, 2018. 1, 2
- [3] Weidong Chen, Dexiang Hong, Yuankai Qi, Zhenjun Han, Shuhui Wang, Laiyun Qing, Qingming Huang, and Guorong Li. Multi-attention network for compressed video referring object segmentation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4416–4425, 2022. 1, 7
- [4] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11030–11039, 2020. 4, 5
- [5] Yi-Wen Chen, Yi-Hsuan Tsai, Tiantian Wang, Yen-Yu Lin, and Ming-Hsuan Yang. Referring expression object segmentation with caption-aware consistency. *arXiv preprint arXiv:1910.04748*, 2019. 1, 2, 3
- [6] 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
- [7] 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. 1, 7
- [8] 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. 6, 7
- [9] 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. 7
- [10] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *European Conference on Computer Vision*, pages 108–124. Springer, 2016. 2
- [11] Guolin Ke, Di He, and Tie-Yan Liu. Rethinking positional encoding in language pre-training. *arXiv preprint arXiv:2006.15595*, 2020. 4
- [12] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 6
- [13] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018. 2
- [14] Dezhuang Li, Ruoqi Li, Lijun Wang, Yifan Wang, Jinqing Qi, Lu Zhang, Ting Liu, Qingquan Xu, and Huchuan Lu. You only infer once: Cross-modal meta-transfer for referring video object segmentation. In *AAAI Conference on Artificial Intelligence*, 2022. 7
- [15] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in Neural Information Processing Systems*, 34, 2021. 1
- [16] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2018. 2
- [17] Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. Dice loss for data-imbalanced nlp tasks. *arXiv preprint arXiv:1911.02855*, 2019. 6
- [18] 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. 2
- [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. 7
- [20] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 5
- [21] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 6
- [22] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 2
- [23] 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. 2
- [24] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*, 2019. 2
- [25] 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. 4
- [26] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 7
- [27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

- [28] Gen Luo, Yiyi Zhou, Rongrong Ji, Xiaoshuai Sun, Jinsong Su, Chia-Wen Lin, and Qi Tian. Cascade grouped attention network for referring expression segmentation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1274–1282, 2020. 2
- [29] 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
- [30] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 1, 2, 6
- [31] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018. 2
- [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. 6
- [33] Hamid Rezaatoughi, 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. 6
- [34] 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, 6, 7
- [35] Hengcan Shi, Hongliang Li, Qingbo Wu, and King Ng Ngan. Query reconstruction network for referring expression image segmentation. *IEEE Transactions on Multimedia*, 23:995–1007, 2020. 3
- [36] Xiaomeng Song, Yucheng Shi, Xin Chen, and Yahong Han. Explore multi-step reasoning in video question answering. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 239–247, 2018. 2
- [37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. Attention is all you need. NIPS’17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. 5
- [38] 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. 5, 6
- [39] 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. 7
- [40] 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. 1, 2, 4, 5, 6, 7
- [41] Sibe Yang, Meng Xia, Guanbin Li, Hong-Yu Zhou, and Yizhou Yu. Bottom-up shift and reasoning for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11266–11275, 2021. 2
- [42] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022. 2
- [43] 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. 7
- [44] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018. 2
- [45] 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. 6
- [46] 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. 4