R^2-VOS: ROBUST REFERRING VIDEO OBJECT SEGMENTATION VIA RELATIONAL CYCLE CONSISTENCY

Anonymous authors
Paper under double-blind review

ABSTRACT

Referring video object segmentation (R-VOS) aims to segment the object masks in a video given a referring linguistic expression to the object. R-VOS introduces human language in the traditional VOS loop to extend flexibility, while all current studies are based on a strict assumption: the object depicted by the expression must exist in the video, namely, the expression and video must have an object-level *semantic consensus*. This is often violated in real-world applications where an expression can be queried to false videos, and existing methods always fail due to abusing the assumption. In this work, we emphasize that studying semantic consensus is necessary to improve the robustness of R-VOS. Accordingly, we pose an extended task from R-VOS without the semantic consensus assumption, named Robust R-VOS (R^2-VOS). The new task essentially corresponds to the joint modeling of the primary R-VOS problem and its dual (text reconstruction). We embrace the observation that, the textual embedding spaces have relational structure consistency in the text-video-text transformation cycle that links the primary and dual problems. We leverage the cycle consistency to consolidate and discriminate the semantic consensus, thus advancing the primary task. We then propose an early grounding module to enable the parallel optimization of the primary and dual problems. To measure the robustness of R-VOS models against unpaired videos and expressions, we construct a new evaluation dataset, R^2-Youtube-VOS. Extensive experiments demonstrate that our method not only identifies negative text-video pairs but also improves the segmentation accuracy for positive pairs with superior disambiguating ability. Our model achieves the state-of-the-art performance on Ref-DAVIS17, Ref-Youtube-VOS, and R^2-Youtube-VOS dataset.

1 INTRODUCTION

Referring video object segmentation (R-VOS) aims to segment a referred object in a video sequence given a linguistic expression. R-VOS has witnessed growing interest thanks to its promising potential in human-computer interaction applications such as video editing and augmented reality. Unlike other video segmentation tasks (Xu et al., 2018; Pont-Tuset et al., 2017) that only rely on visual cues, R-VOS (Khoreva et al., 2018) pairs a target video with a linguistic expression referring to an object.

Previous works (Botach et al., 2021; Wu et al., 2022) tackle the R-VOS problem with a strict assumption that the referred object exists in the video, i.e., there is an object-level semantic consensus between the expression and the video. However, this assumption does not always hold in practice. As shown in the second row of Figure 1, we notice a severe false-alarm problem experienced by previous methods when the semantic consensus does not exist, blocking such methods in various applications that cannot provide accurate vision-language pairs. We argue that the current R-VOS task is not completely defined with the assumption that the referred object always exists in the video.

Even when semantic consensus exists in the video-language pairs, it is still challenging to locate the correct object due to the multimodal nature of the R-VOS task. Recently, MTTR (Botach et al., 2021) employs a multimodal transformer encoder to learn a joint representation of the linguistic expression and video, and then obtains the referred object by ranking all objects in the video. ReferFormer (Wu et al., 2022) follows the image-level method, ReTR (Li & Sigal, 2021), to adopt the linguistic expression as a query to the transformer decoder to avoid redundant ranking of all objects. However, these latest methods suffer from semantic misalignment of the segmented object and the linguistic expression, even with sophisticated components employed. As shown in the first row of Figure 1, the segmented objects by MTTR and ReferFormer are not the object referred to.

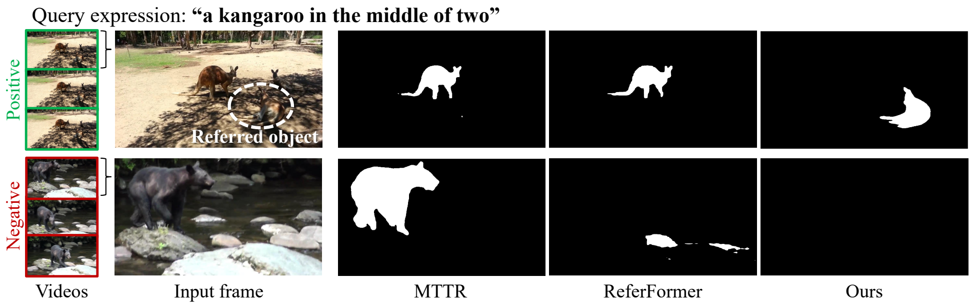

Figure 1: Illustration of the new R^2-VOS task. A linguistic expression is given to query a set of videos without the semantic consensus assumption. Videos containing the referred object by the expression are positive, otherwise negative. Unlike the previous R-VOS setting that assumes all target videos are positive to the query expression, the new R^2-VOS task is required to discriminate positive and negative text-video pairs, and further segment object masks in positive videos or treat entire negative videos as backgrounds. Compared to the previous R-VOS methods, MTTR (Botach et al., 2021) and ReferFormer (Wu et al., 2022), our method not only discriminates negative videos better but also shows a superior disambiguating ability between visually similar objects in positive videos.

In this paper, we seek to investigate the semantic alignment problem between visual and linguistic modalities in referring video segmentation. We extend the current task definition of R-VOS (Khoreva et al., 2018) to accept both paired and unpaired video-language inputs. This new task, which we term Robust R-VOS (R^2-VOS), overcomes the limitation of the R-VOS task by additionally considering the semantic alignment of input videos and referring expressions. We reveal that this task is essentially related to two problems that are interrelated (Mao et al., 2016): the R-VOS problem as the **primary** problem of segmenting mask sequences from videos with referring texts, and its **dual** problem of reconstructing text expressions from videos with object masks. By linking the primary and dual problems, we introduce a text-video-text cycle and a corresponding relational consistency constraint. This cycle constraint can 1) improve the segmentation accuracy by enforcing the semantic consensus between paired text query and segmented mask, and 2) discriminate semantic misalignment by assessing an explicit cycle consistency criteria to alleviate the false alarm problem. Although there are previous works (Shi et al., 2020; Chen et al., 2019) on referring image segmentation utilizing cyclic training, the primary segmentation task could be degraded due to the improper dual problem, because they try to reconstruct deterministic text expressions while the pretrained linguistic model has dataset bias for expressions. Differently, our cycle constraint is applied to the textual embedding space, which circumvents the raw dataset bias problem. Specifically, we equip the cycle with an early grounding module, which can handle the primary-dual tasks in a parallel manner and also can manipulate a relational cyclic constraint to preserve the structures between the input and reconstructed textural embedding spaces. In addition, the early grounding module benefits to locate the correct object by suppressing irrelevant features in an early stage. Our contributions can be summarized as:

- We are the first to address the severe false-alarm problem faced by previous R-VOS methods with unpaired video-text inputs. To investigate the robustness of R-VOS models, we introduce the new R^2-VOS task accepting unpaired inputs, as well as an evaluation dataset and corresponding metrics.
- We introduce a relational cycle consistency constraint to consolidate the semantic alignment between visual and textual modalities, and also discriminate false-positive by assessing the cycle consistency criteria.
- We propose a novel early grounding module to locate the referred object in an early stage, serving as a proxy, to bridge the primary referring segmentation and dual expression reconstruction task for joint optimization.
- Our method outperforms previous state-of-the-art methods on Ref-Youtube-VOS, Ref-DAVIS, and R^2-Youtube-VOS dataset.

2 Related Works

Vision and language representation learning. There have been a long line of studies on how to learn vision-language representation, e.g., multimodal attention (Luo et al., 2020; Gao et al., 2019), fusion scheme (Fukui et al., 2016; Kim et al., 2018), multi-step reasoning (Yang et al., 2016; Hudson & Manning, 2018) and pretraining (Radford et al., 2021). KAC Net (Chen et al., 2018) leverages knowledge-aided consistency constraints to enhance semantic alignment for weakly supervised phrase grounding. A structure-preserving constraint (Wang et al., 2016) is proposed to preserve some intra-modal properties when learning vision-language representation for image-text retrieval.

Referring video object segmentation. URVOS (Seo et al., 2020) is the first unified R-VOS framework with a cross-modal attention and a memory attention module, which largely improves R-VOS performance. ClawCraneNet (Liang et al., 2021a) leverages cross-modal attention to bridge the semantic correlation between textual and visual modalities. ReferFormer (Wu et al., 2022) and MTTR (Botach et al., 2021) are two latest works that utilize transformers to decode or fuse multimodal features. ReferFormer (Wu et al., 2022) employs a linguistic prior to the transformer decoder to focus on the referred object. MTTR (Botach et al., 2021) leverages a multimodal transformer encoder to fuse linguistic and visual features. Different from other vision-language tasks, e.g., image-text retrieval (Lin et al., 2014; Liu et al., 2019a; Miech et al., 2018) and video question answering (Lei et al., 2018; Song et al., 2018), R-VOS needs to construct object-level multimodal semantic consensus in a dense visual representation.

3 R^2-VOS

Task definition. We introduce a new 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 previous R-VOS setup, the queried videos are not required to contain the referred object by expression E_o. A video V and an expression E_o have **semantic consensus** if the object o appears in V, and the video is **positive** with respect to E_o, otherwise it is **negative**. The R^2-VOS task is extended to discriminate positive and negative videos, and predict masks M_o of object o for positive videos and treat all frames in the negative videos as background.

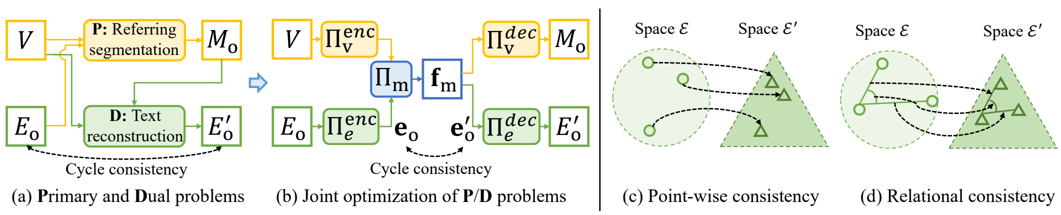

Figure 2: Problem analysis. (a) R^2-VOS introduces the **P**rimary problem of referring segmentation and the **D**ual problem of text reconstruction for positive videos. The **P/D** problems are connected in a cycle path from original expression E_o to reconstructed expression E'_o. (b) The cycle consistency between the original and reconstructed embeddings ($\mathbf{e}_o$ and $\mathbf{e}'_o$) can benefit to optimize the **P** problem. We enable the joint optimization for cycle consistency with a cross-modal proxy $\mathbf{f}_m$ defined between all single-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 because the mapping between $\mathcal{E}$ and $\mathcal{E}'$ are not necessarily bijective. An object can be referred by various textual expressions. (d) Instead, we apply a relational consistency to preserve distances and angles.

Primary and dual problems for R^2-VOS. 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:

$$p(M_o|V, E_o) \sim p(E_o|V, M_o)p(M_o|V) \tag{1}$$

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. We note that for negative videos, $p(E_o|V, M_o)$ is undefined because the mask M_o is empty. Thus, we only investigate the dual problem for positive videos. The primary problem and the dual problem can be connected in a cycle path, i.e., from the original expression E_o to the reconstructed expression E'_o through positive video queries, as shown in Figure 2 (a). We believe that the cycle constraint benefits to optimize the primary problem by enhancing the semantic consensus.

In practice, we study the cycle consistency between the original textual embedding space $\mathcal{E}$ and the transformed textual embedding space $\mathcal{E}'$ induced by positive videos. By definition, the path from the original text embedding $\mathbf{e}_o$ to the reconstructed text embedding $\mathbf{e}'_o$ is modulated with **cross-modal** interactions between video and text. Thus, to link the primary and dual problem and enable the joint

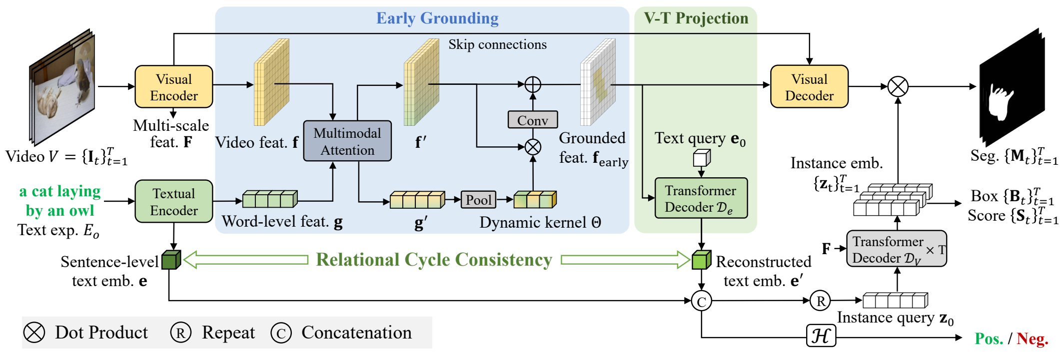

Figure 3: Overview of the proposed model. Given a video clip $V = \{\mathbf{I}_t\}_{t=1}^{T}$ and a textual expression E_o referring object o, we first extract video feature and text feature separately, then fuse them in the early grounding module to obtain the visual representation $\mathbf{f}_{\mathbf{early}}$ of the referred object o. Then we project $\mathbf{f}_{\mathbf{early}}$ to a textual space to be $\mathbf{e}'$ and add the relational cycle constraint with the original text embedding $\mathbf{e}$. The final segmentation is obtained by dynamic convolutions with video features from the visual decoder and dynamic weights from the fused text embeddings. The semantic consensus of input pairs is discriminated to be positive or negative by assessing the consistency between $\mathbf{e}$ and $\mathbf{e}'$.

optimization, we introduce a cross-modal intermediate feature $\mathbf{f}_m$ to convey information of both the input of the primary problem (V, E_o) and the dual problem (V, M_o), as shown in Figure 2 (b). $\mathbf{f}_m$ is defined between the encoder and decoder stages of single-modal operations, i.e., Π_v^{enc}, Π_e^{enc}, Π_v^{dec}, Π_e^{dec}, to only focus on the multi-modal interaction.

Relational cycle consistency. A key observation for cycle consistency between $\mathcal{E}$ and $\mathcal{E}'$ is that the mapping between them is not necessarily bijective, as there could be multiple textual descriptions for the same object. Thus, naively adding point-wise consistency, i.e., $\mathbf{e}_o = \mathbf{e}'_o, \forall \mathbf{e}_o \in \mathcal{E}$ will collapse the feature space to a sub-optimal solution. Instead, we take inspiration from relational knowledge distillation (Park et al., 2019), and introduce relational cycle consistency for $\mathcal{E}$ and $\mathcal{E}'$. The relational cycle consistency is to preserve the structure of the feature space rather than exact point-wise consistency, as illustrated in Figure 2 (c) and (d). Mathematically, the structure-preserving property is defined as isometric and conformal constraints to preserve pair-wise distance and angles for $\mathbf{e} \in \mathcal{E}$ and $\mathbf{e}' \in \mathcal{E}'$:

$$|\mathbf{e}_1 - \mathbf{e}_2| = |\mathbf{e}'_1 - \mathbf{e}'_2| \quad (2)$$

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

where $|\cdot|$ and $\angle(\cdot)$ denote distance and angle metrics.

4 METHOD

In this section, we elaborate our R^2-VOS framework with the relational consistency, which mainly consists of four parts: feature extraction, early grounding as a proxy, video-text (V-T) projection for text reconstruction, and mask decoding for final segmentation, as shown in Figure 3. We first extract the video feature $\mathbf{f}$, word-level text feature $\mathbf{g}$, and sentence-level text embedding $\mathbf{e}$. On the one hand, to model the primary segmentation problem of maximizing $p(M_o|V, E_o)$, we enable the multimodal interaction in the early grounding module to generate the grounded feature $\mathbf{f}_{\mathbf{early}}$. $\mathbf{f}_{\mathbf{early}}$ coarsely locates the referred object o and filters out irrelevant features, which serves as a proxy linking the primary segmentation and dual text reconstruction problem. The final mask M_o is obtained by dynamic convolution (Chen et al., 2020) on the decoded visual feature maps, with kernels learned from instance embedding $\{\mathbf{z}_{\mathrm{t}}\}_{t=1}^{T}$. On the other hand, to model the dual text reconstruction problem of maximizing $p(E_o|V, M_o)$, we utilize the grounded video feature $\mathbf{f}_{\mathbf{early}}$ as the alternative of V and M_o, since $\mathbf{f}_{\mathbf{early}}$ conveys contextual video clues of object o. In this way, we enable the parallel optimization of the primary and dual problem by relating them to $\mathbf{f}_{\mathbf{early}}$. Specifically, we employ a V-T projection module to project $\mathbf{f}_{\mathbf{early}}$ onto a reconstructed text embedding $\mathbf{e}'$. We add a relational 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 semantic consensus discrimination head $\mathcal{H}(\mathbf{e}, \mathbf{e}')$ to assess the consistency between original and reconstructed text embeddings, discriminating the alignment of multimodal semantics and identifying negative videos.

4.1 Single-modal Feature Extraction

Visual encoder. Following previous methods (Botach et al., 2021; Wu et al., 2022; Wang et al., 2021), we build the visual encoder with a visual backbone and a deformable transformer encoder (Zhu et al., 2020) on top of it. The extracted features from the backbone are flattened, projected to a lower dimension, added with positional encoding (Ke et al., 2020), and then fed into a deformable transformer encoder (Zhu et al., 2020) similar to the previous method (Wu et al., 2022). 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 (Liu et al., 2019b) 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 Early Grounding

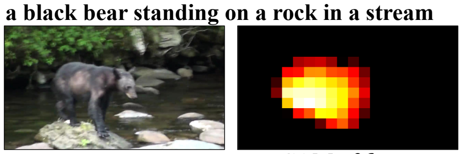

Figure 4: Visualization of channel activation map (CAM) of $\mathbf{f_{early}}$.

We propose an early grounding module to coarsely locate the referred object o and filter out irrelevant features. Then the grounded feature $\mathbf{f_{early}}$ 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 connecting the two problems. Figure 4 shows a visualization of $\mathrm{f_{early}}$. Specifically, we utilize the power of dynamic convolution (Chen et al., 2020) to discriminate visual features in the early stage. As shown in the blue part of Figure 3, we first enable the multimodal interaction between video and text features, then apply the dynamic convolution with kernels learned from text feature to discriminate the object-level semantics. In particular, multi-head cross-attention (MCA) (Vaswani et al., 2017) is leveraged to conduct the multimodal interaction:

$$\mathbf{h_f} = \mathrm{LN}(\mathrm{MCA}(\mathbf{f}, \mathbf{g}) + \mathbf{f}) \quad \mathbf{f}' = \mathrm{LN}(\mathrm{FFN}(\mathbf{h_f}) + \mathbf{h_f}) \tag{4}$$

$$\mathbf{h_g} = \mathrm{LN}(\mathrm{MCA}(\mathbf{g}, \mathbf{f}) + \mathbf{g}) \quad \mathbf{g}' = \mathrm{LN}(\mathrm{FFN}(\mathbf{h_g}) + \mathbf{h_g}), \tag{5}$$

where $\mathrm{MCA}(\mathbf{X}, \mathbf{Y}) = \mathrm{Attention}(\mathbf{W^Q X}, \mathbf{W^K Y}, \mathbf{W^V Y})$. $\mathbf{W}$ represents learnable weight. LN and FFN denote layer normalization and feed-forward network respectively. The text feature $\mathbf{g}'$ is further pooled to a fixed length, and followed by a fully-connected layer to form the dynamic 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 $\mathbf{f_{early}} \in \mathbb{R}^{C \times THW}$

$$\mathbf{f_{early}} = \mathrm{BN}(\varphi(\theta_1 \mathbf{f}' \oplus \cdots \oplus \theta_K \mathbf{f}') + \mathbf{f}'), \tag{6}$$

where $\oplus$ is the concatenation in channel dimension and $\varphi(\cdot)$ is a convolution to reduce the feature dimension. BN denotes batch normalization.

4.3 Text Reconstruction

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

4.4 Referring Segmentation

Mask segmentation. Similar to previous methods (Wu et al., 2022; Botach et al., 2021; Kamath et al., 2021), we leverage deformable transformer decoders with dynamic convolution to segment the object masks. 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 Figure 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 (Wang et al., 2021) $\mathbf{z}_0 \in \mathbb{R}^{C_q \times N}$, 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 $\mathbf{f}_{\mathbf{out},t} \in \mathbb{R}^{C\times H\times W}$ is obtained by fusing low-level features from the feature pyramid network (Lin et al., 2017a) in the visual decoder. The mask prediction $\mathbf{M}_t \in \mathbb{R}^{N\times H\times W}$ can be computed by $\mathbf{M}_t = \mathbf{w}_t^{\mathrm{T}} \mathbf{f}_{\mathbf{out},\mathrm{t}}$.

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:

$$\mathcal{L} = \lambda_{text}\mathcal{L}_{text} + \lambda_{segm}\mathcal{L}_{segm} + \lambda_{align}\mathcal{L}_{align}, \tag{7}$$

where $\mathcal{L}_{text}$, $\mathcal{L}_{segm}$, and $\mathcal{L}_{align}$ are losses for text reconstruction, referring segmentation and semantic consensus discrimination respectively. A ground-truth semantic alignment $\hat{A} = \{0, 1\}$ is utilized to discriminate positive and negative pairs. The $\mathcal{L}_{align}$ is simply a cross-entropy loss between the predicted alignment A and ground-truth $\hat{A}$. The other two terms are computed as follows:

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

$$\mathcal{L}_{text} = \mathbb{1}(\hat{A}) \cdot (\mathcal{L}_{dist} + \lambda_{angle}\mathcal{L}_{angle}), \tag{8}$$

where the indicator function $\mathbb{1}(\hat{A}) = 1$ if the alignment indicates the referred object exists in the video, otherwise 0, λ_{angle} is a hyperparameter balancing the distance loss $\mathcal{L}_{dist}$ and angle loss $\mathcal{L}_{angle}$. We elaborate these two losses according to the relational cycle consistency Equation 2. Let $\mathcal{X}^n = \{(x_1, ..., x_n)|x_i \in \mathcal{X}\}$ denote a set of n-tuples, $\Phi^n = \{(\mathbf{x}, \mathbf{x}')|\mathbf{x} \in \mathcal{X}^n, \mathbf{x}' \in \mathcal{X}'^n\}$ 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:

$$\mathcal{L}_{dist} = \sum_{(\mathbf{x}, \mathbf{x}')\in\Phi^2} l_\delta(\phi_D(\mathbf{x}), \phi_D(\mathbf{x}')), \quad \phi_D(\mathbf{x}) = \tfrac{1}{\mu(\mathbf{x})}\|\mathbf{e}_i - \mathbf{e}_j\|_2, \tag{9}$$

$$\mathcal{L}_{angle} = \sum_{(\mathbf{x}, \mathbf{x}')\in\Phi^3} l_\delta(\phi_\angle(\mathbf{x}), \phi_\angle(\mathbf{x}')), \quad \phi_\angle(\mathbf{x}) = \cos\angle(\mathbf{e}_i, \mathbf{e}_j, \mathbf{e}_k), \tag{10}$$

where $\mu(\mathbf{x}) = \sum_{\mathbf{x}=(x_1,x_2)\in\mathcal{X}^2} \frac{||x_1-x_2||_2}{|\mathcal{X}^2|}$ is the average distance function, and 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 $\emptyset$). The similarity can be computed as

$$\mathcal{L}_{match}(\mathbf{y}_i, \hat{\mathbf{y}}) = \lambda_{box}\mathcal{L}_{box} + \lambda_{conf}\mathcal{L}_{conf} + \lambda_{mask}\mathcal{L}_{mask}, \tag{11}$$

where λ_{box}, λ_{conf}, and λ_{mask} are weights to balance losses. Following previous works (Ding et al., 2021; Wang et al., 2021), we leverage a combination of Dice (Li et al., 2019) and BCE loss as $\mathcal{L}_{mask}$, focal loss (Lin et al., 2017b) as $\mathcal{L}_{conf}$, and GIoU (Rezatofighi et al., 2019) and L1 loss as $\mathcal{L}_{box}$. The best assignment $\hat{\sigma}$ is solved by Hungarian algorithm (Kuhn, 1955). 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:

$$\{\bar{\mathbf{M}}_t\}_{t=1}^{T} = \{\mathbb{1}(A > 0.5) \cdot \mathbf{M}_{\bar{s},t}\}_{t=1}^{T}, \quad \bar{\mathbf{s}} = \arg\max_i \{\mathbf{S}_{i,1} + \cdots + \mathbf{S}_{i,T}\}_{i=1}^{N}, \tag{12}$$

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 $\mathbf{M}_t$ respectively. $\bar{\mathbf{s}}$ 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 alarm. $\mathbb{1}(A > 0.5) = 1$ if $A > 0.5$, else 0.

5 Experiment

5.1 Dataset and Metrics

Dataset. We conduct experiments on three datasets: Ref-Youtube-VOS, Ref-DAVIS and R^2-Youtube-VOS. Ref-Youtube-VOS (Seo et al., 2020) is a large-scale benchmark that has 3,978 videos with about $15k$ language descriptions. There are 3,471 videos with 12,913 expressions in the training set and 507 videos with 2,096 expressions in the validation set. Ref-DAVIS-17 (Khoreva et al., 2018) contains 90 videos with 1,544 expressions, including 60 and 30 videos for training and validation respectively. We construct a new **evaluation** dataset, R^2-Youtube-VOS, which extends the Ref-Youtube-VOS validation set with each expression querying two videos, a positive video (the same in Ref-Youtube-VOS) and a negative video. The negative text-video pairs are constructed by shuffling the original ordered videos and constraining all expressions and videos unmatched. The segmentation accuracy is evaluated on the positive text-video pairs, thus the same as on Ref-Youtube-VOS. In the training, we use the original Ref-Youtube-VOS training set, but we randomly pick unmatched text-video pairs as negative samples as augmentation.

Metrics. We employ commonly-used region similarity $\mathcal{J}$ and contour accuracy $\mathcal{F}$ (Pont-Tuset et al., 2017) for conventional Ref-Youtube-VOS and Ref-DAVIS-17 benchmarks. For the proposed R^2-Youtube-VOS task, we additionally introduce a 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 object false alarm 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 (Ding et al., 2021; Wu et al., 2022), our model is first pre-trained on Ref-COCO/+/g dataset (Yu et al., 2016; Mao et al., 2016) 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 $\mathrm{batchsize}$ of 8 and an AdamW (Loshchilov & Hutter, 2017) optimizer with weight decay 1×10^{-4}. Following convention (Botach et al., 2021), 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 5 for all backbones. We create negative pairs by shuffling positive pairs in each batch. Our method is implemented with PyTorch (Paszke et al., 2019). More details can be found in Appendix.C.

5.3 Main Results

We compare our method with state-of-the-art R-VOS methods on Ref-Youtube-VOS and Ref-DAVIS-17, and R^2-VOS task in Table 1. **Comparison on Ref-Youtube-VOS.** In Table 1, we first compare our method on Ref-Youtube-VOS. For results of ResNet (He et al., 2016) backbone, our method achieves 57.3 $\mathcal{J}\&\mathcal{F}$ which outperforms the latest method ReferFormer (Wu et al., 2022) 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 single NVIDIA P40 with $batchsize = 1$). For results of Swin-Transformer (Liu et al., 2021) 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 (Wu et al., 2022) and MTTR (Botach et al., 2021) by a clear margin. More analysis is available in the Appendix B.1.

Method	Backbone	Ref-Youtube-VOS			R^2-Youtube-VOS	Ref-DAVIS-17		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{R}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Spatial Visual Backbone								
CMSA (Ye et al., 2019)	ResNet-50	34.9	33.3	36.5	-	34.7	32.2	37.2
CMSA + RNN (Ye et al., 2019)	ResNet-50	36.4	34.8	38.1	-	40.2	36.9	43.5
URVOS (Seo et al., 2020)	ResNet-50	47.2	45.3	49.2	-	51.5	47.3	56.0
PMINet (Ding et al., 2021)	ResNet-101	53.0	51.5	54.5	-	-	-	-
CITD (Liang et al., 2021b)	ResNet-101	56.4	54.8	58.1	-	-	-	-
ReferFormer (Wu et al., 2022)	ResNet-50	55.6	54.8	56.5	30.6	58.5	55.8	61.3
Ours	ResNet-50	**57.3**	**56.1**	**58.4**	**94.1**	**59.7**	**57.2**	**62.4**
ReferFormer (Wu et al., 2022)	Swin-T	58.7	57.6	59.9	28.2	-	-	-
Ours	Swin-T	**60.2**	**58.9**	**61.5**	**94.4**	-	-	-
Spatio-temporal Visual Backbone								
MTTR (Botach et al., 2021)	Video-Swin-T	55.3	54.0	56.6	5.9	-	-	-
ReferFormer (Wu et al., 2022)	Video-Swin-T	59.4	58.0	60.9	28.5	-	-	-
Ours	Video-Swin-T	**61.3**	**59.6**	**63.1**	**95.7**	-	-	-

Table 1: **Comparison to state-of-the-art R-VOS methods.**

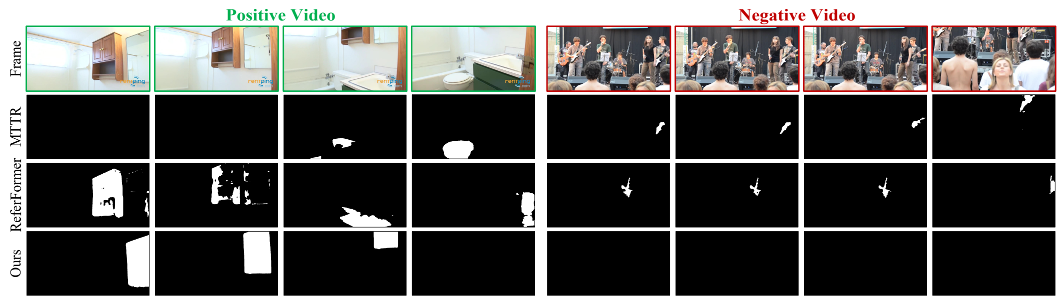

Figure 6: Qualitative comparison to the state-of-the-art R-VOS method on the R^2-VOS task.

Comparison on Ref-DAVIS-17. Our method achieves 59.7 $\mathcal{J}\&\mathcal{F}$ on Ref-DAVIS-17 dataset, which outperforms ReferFormer by 1.2 $\mathcal{J}\&\mathcal{F}$. **Comparison on R^2-VOS.** As shown in Table 1, the state-of-the-art R-VOS methods, ReferFormer and MTTR suffer from a low $\mathcal{R}$ metric which measures the false-alarm problem when the semantic consensus of the input text-video pair does not hold. Compared to the severe false alarm of previous R-VOS methods, our model successfully mitigates the false alarm of the model, thanks to the proposed multimodal cycle consistency constraint and semantic consensus discrimination.

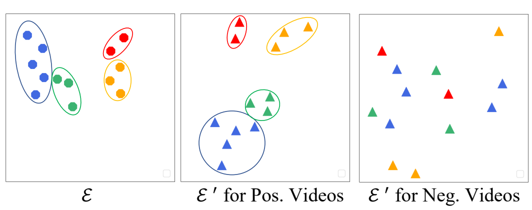

Figure 5: Visualization of text embedding spaces. Dots represent original text embeddings in $\mathcal{E}$, and triangles represent reconstructed ones in $\mathcal{E}'$ induced by positive and negative videos respectively. Elements in the same color belong to the same object. Note that an object can 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.

Qualitative results. We compare the qualitative results of our method against state-of-the-art methods in Figure 6 on R^2-VOS. For **positive videos**: The result indicates that our method predicts accurate and temporally-consistent results, while ReferFormer (Wu et al., 2022) and MTTR (Botach et al., 2021) fail to locate the correct object. For **negative videos**: Both ReferFormer and 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 the consensus discrimination works, we visualize the text embedding and reconstructed text embedding spaces for both positive and negative videos. As shown in Figure 5, 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. To investigate the effectiveness of different components in our method, we conduct experiments with the ResNet-50 backbone on R^2-Youtube-VOS dataset. We build a transformer-based baseline model and equip our proposed components step-by-step. As shown in

Components	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{R}$
Baseline	52.4	51.9	52.8	34.9
+EG	55.5$_{+3.1}$	54.4	56.5	32.9$_{-2.0}$
+EG+FT	55.5$_{+3.1}$	54.5	56.5	33.4$_{-1.5}$
+EG+CC	56.9$_{+4.5}$	55.7	58.1	94.0$_{+59.1}$
+EG+CC+FT	**57.3**$_{+4.9}$	**56.1**	**58.4**	**94.1**$_{+59.2}$

Table 2: **Impact of different components in our method.** EG: Early grounding, CC: Consistency constraint, FT: Fusing text embeddings.

Method	NS	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{R}$
ReferFormer	✗	55.6	54.8	56.5	30.6
ReferFormer*	✓	42.2	41.2	43.2	63.3
Ours	✗	57.2	**56.1**	58.3	46.8
Ours	✓	**57.3**	**56.1**	**58.4**	**94.1**

Table 4: **Impact of the negative samples.**

Constraint	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{R}$
None	55.6	54.6	56.5	66.3
PW	54.4$_{-1.2}$	53.3	55.5	88.7$_{+22.4}$
RA	56.7$_{+1.1}$	55.5	57.9	93.6$_{+27.3}$
RD	56.4$_{+0.8}$	55.2	57.6	90.4$_{+24.1}$
RD+RA	**56.9**$_{+1.3}$	**55.7**	**58.1**	**94.0**$_{+27.7}$

Table 3: **Impact of the cycle consistency constraint.** PW: Point-wise. RA: Relational angle. RD: Relational distance.

Query Number	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{R}$
1	54.9	54.2	55.6	**94.7**
5	**57.3**	56.1	**58.4**	94.1
9	57.0	**56.8**	57.2	93.5

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

Table 2, the baseline model achieves 52.4 $\mathcal{J}\&\mathcal{F}$. After employing the early grounding module, the performance boosts to 55.5 $\mathcal{J}\&\mathcal{F}$ and the cycle-consistency constraint with negative training samples brings another 1.4 $\mathcal{J}\&\mathcal{F}$ gain. By using the fused text embedding as instance query, we achieve our best performance of 57.3 $\mathcal{J}\&\mathcal{F}$.

Consistency constraint. We conduct experiments to ablate the influence of cycle-consistency constraints. As shown in Table 3, utilizing point-wise consistency constraint will lead to a performance drop compared to the setting without cycle 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 early grounding 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 $\mathbf{f}_{\mathbf{early}}$ and result in a collapse for text reconstruction. Table 3 shows that sole relational angle constraint can bring 1.2 $\mathcal{J}\&\mathcal{F}$ gain, and it can be slightly improved with 1.4 $\mathcal{J}\&\mathcal{F}$ gain by jointly using relational angle and distance constraint.

Negative training samples for discrimination head. To study the effects of introducing negative samples in the training on different pipelines, we augment the original ReferFormer as ReferFormer* with an additional classification head after the text query and visual FPN (the same head for predicting reference score in Section 3.4 of the paper (Wu et al., 2022)) to discriminate negative videos. Training with the same data (containing positive and negative samples), we notice that ReferFormer* does not achieve comparable results as ours, and is even worse than its original version with only positive training samples, as shown in Table 4. Negative training samples may degrade the segmentation quality since they only predict blank masks. Naively adding a classification head does not work well. The reasons that our method can fully utilize negative samples to improve model robustness could be that 1) our discrimination head $\mathcal{H}$ is based on the cycle consistency, which straightforwardly expresses the degree of alignment between visual and textual modalities, 2) $\mathcal{H}$ affect the visual decoder less in our pipeline as shown in Figure 3. More analysis is available in Appendix.B.

Instance query number. Although only one referral is involved for each frame in R-VOS task, 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 result.

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 relational cycle consistency constraint, is introduced to discriminate and enhance the semantic consensus between visual and textual modalities. To evaluate the model robustness, we extend the R-VOS task to accept unpaired inputs and collect a corresponding R^2-Youtube-VOS dataset. We observe a severe false-alarm problem suffered from previous methods on R^2-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 Ref-DAVIS17, Ref-Youtube-VOS, and R^2-VOS dataset. We believe that, with unpaired inputs, R^2-VOS is a more general setting of referring video segmentation, which can shed light on a new direction to investigate the robustness of referring segmentation.

REFERENCES

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

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*, pp. 213–229. Springer, 2020.

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*, pp. 4042–4050, 2018.

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.

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*, pp. 11030–11039, 2020.

Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. *arXiv preprint arXiv:2104.08541*, 2021.

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*, pp. 7, 2021.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6639–6648, 2019.

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*, pp. 770–778, 2016.

Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067*, 2018.

Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers. *arXiv preprint arXiv:2106.03299*, 2021.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1780–1790, 2021.

Guolin Ke, Di He, and Tie-Yan Liu. Rethinking positional encoding in language pre-training. *arXiv preprint arXiv:2006.15595*, 2020.

Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Asian Conference on Computer Vision*, pp. 123–141. Springer, 2018.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *Advances in neural information processing systems*, 31, 2018.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018.

Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in Neural Information Processing Systems*, 34, 2021.

Xiang Li, Jinglu Wang, Xiao Li, and Yan Lu. Hybrid instance-aware temporal fusion for online video instance segmentation. *arXiv preprint arXiv:2112.01695*, 2021.

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.

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

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*, 2021b.

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*, pp. 740–755. Springer, 2014.

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*, pp. 2117–2125, 2017a.

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*, pp. 2980–2988, 2017b.

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*, 2019a.

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*, 2019b.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020.

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*, pp. 11–20, 2016.

Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

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*, pp. 658–666, 2019.

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*, pp. 208–223. Springer, 2020.

Hengcan Shi, Hongliang Li, Qingbo Wu, and King Ngi Ngan. Query reconstruction network for referring expression image segmentation. *IEEE Transactions on Multimedia*, 23:995–1007, 2020.

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*, pp. 239–247, 2018.

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, pp. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5005–5013, 2016.

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*, pp. 8741–8750, 2021.

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

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

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 21–29, 2016.

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*, pp. 10502–10511, 2019.

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

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.

Appendix

A ADDITIONAL EXPERIMENTS

A.1 FRAME NUMBER

Window Size	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{R}$
1	53.5	53.0	54.0	89.2
3	56.8	**56.5**	57.1	92.1
5	**57.3**	56.1	**58.4**	**94.1**

Table A: **Impact of the window size.**

Since R-VOS gives a text that describes an object over a period of time, temporal information is vital to segment accurate and temporally-consistent results. We ablate on the best window size of input videos during training. As shown in Table A, we notice that the performance improves as the window size increases and a window size of 5 brings the best result of 57.3 $\mathcal{J}\&\mathcal{F}$.

A.2 NEGATIVE VIDEOS WITHOUT POSITIVE TEXT

Negative Video Source	$\mathcal{R}$	
	ReferFormer	Ours
Ref-Youtube-VOS	30.6	94.1
Ref-Youtube-VOS & Ref-DAVIS	33.1	92.2

Table B: Impact of different negative video sources.

As shown in Table B, we test the robustness of our model on two settings. We generate negative videos from Ref-Youtube-VOS and a combination of Ref-Youtube-VOS and Ref-DAVIS dataset. In both settings, all videos in the validation set are leveraged. The results indicates that source of negative videos has minor impact on the robustness of our model.

A.3 DYNAMIC KERNEL NUMBER IN EARLY GROUNDING MODULE

L_θ	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
1	56.4	55.1	57.7
2	57.0	55.3	58.0
3	**57.3**	**56.1**	**58.4**
4	57.1	55.9	58.2

Table C: **Impact of the dynamic filter number.**

As shown in Table C, we conduct experiments to investigate the impact of the dynamic filter number in the early grounding module. The dynamic convolution is extensively used to decode dense features in video instance segmentation (Hwang et al., 2021; Li et al., 2021) and object detection (Carion et al., 2020) because of its strong ability to generate instance-specific filters to modify the feature maps. In our method, we use a text-guided dynamic convolution to ground referred object in the feature level. We notice that using a dynamic kernel number of 3 brings the best performance.

A.4 SEMANTIC ALIGNMENT DISCRIMINATION

As shown in Table D, we conduct experiments without using the semantic alignment $\mathbb{1}(A)$ to filter out negative videos during inference. We notice that, even if $\mathbb{1}(A)$ is not applied to the final output, our model has a much higher $\mathcal{R}$ score compared to previous methods on R^2-Youtube-VOS. This indicates the consistency constraint can boost the model robustness to negative videos even without explicitly filtering out videos with semantic alignment discrimination.

Method	Backbone	$\mathcal{J}\&\mathcal{F}$ & $\mathcal{R}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{R}$
ReferFormer (Wu et al., 2022)	ResNet-50	47.3	54.8	56.5	30.6
Ours	ResNet-50	**59.2**	**56.1**	**58.4**	**63.2**
MTTR (Botach et al., 2021)	Video-Swin-T	40.0	55.9	58.1	5.9
ReferFormer (Wu et al., 2022)	Video-Swin-T	49.1	58.0	60.9	28.5
Ours	Video-Swin-T	**62.7**	**59.6**	**63.1**	**65.5**

Table D: **Comparison to state-of-the-art R-VOS methods on R^2-Youtube-VOS without applying $\mathbb{1}(A)$ to filter out videos during inference.**

B More Quantitative Result Analysis

B.1 Performance Gain Analysis

Under the same ResNet-50 backbone, our method achieves 57.3 $\mathcal{J}\&\mathcal{F}$, 94.1 $\mathcal{R}$ and 30 FPS compared to the 55.6 $\mathcal{J}\&\mathcal{F}$, 30.6 $\mathcal{R}$ and 22 FPS of ReferFormer. We will then point-to-point analyze reasons of improvements on $\mathcal{J}\&\mathcal{F}$ (for positive video), $\mathcal{R}$ (for negative videos) and FPS (for inference speed).

- $\mathcal{J}\&\mathcal{F}$: (1) We introduce the early-grounding module which employs both pixel-wise and channel-wise attention to enable multimodal interaction. Different from the CM-FPN used in ReferFormer that solely fuses features from text to video in pixel-level, our early-grounding module first enables pixel-level bi-directional fusion and then generates dynamic kernels using the fused text feature $\mathbf{g}'$ to modulate the video feature $\mathbf{f}'$. The dynamic convolution (channel-wise attention) is commonly used to decode dense masks from visual features and is suitable to suppress irrelevant features. By equipping text-guided dynamic convolution in early-stage, the pixel decoder can be more focused on the target object (as shown in Figure 4). (2) Our method leverages relational cycle consistency to constraint the intermediate feature $\mathrm{f}_{\mathrm{early}}$ to contain correct object-level information to recover some properties of original text embedding. By applying this constraint, our method can better avoid interference and easier locate the correct object. (3) Our instance query is composed of both original sentence embedding and the reconstructed one. Different from ReferFormer that only utilizes original sentence embedding as queries, the reconstructed embedding can encode visual information to facilitate the instance query decode the objects from visual features.
- $\mathcal{R}$: The newly introduced metric $\mathcal{R}$ aims to measure the robustness of the model against unpaired inputs. Text-video pairs with (object-level) semantic consensus can be assumed as in-distribution for RVOS problem where semantic consensus can be kind of easily modeled. In contrast, unpaired text-video is much more difficult to tackle because there can be unlimited out-of-distribution (OOD) scenarios for the text-video pairs. In our method, instead of directly detect the OOD of input pairs, we convert the problem to find semantic alignment between the input text embedding and reconstructed embedding and constraint the property of reconstructed space by introducing the cycle consistency. In this way, the comparison is conducted in the constraint original and reconstructed text spaces. For ReferFormer, it models the alignment of text to video by querying the visual features by text in the transformer decoder. In this way, the comparison is conducted in unconstrained text and video spaces thus results in a inferior performance.
- FPS: The speed improvement of our method mainly comes from our efficient multimodal fusion. Compared to the multi-scale CM-FPN, our early-grounding module is only conduct at the high-level. In addition, our bi-direction multimodal fusion (Equ 4 & 5) only leverages cross-attention to avoid computational expensive video-to-video operations.

B.2 Failure of ReferFormer with Negative Training Samples

Adding a background class to ReferFormer and training with negative samples cannot benefit ReferFormer. The principal difference between that and our method is that between implicit and explicit classes. In the absence of negative samples, a "none of the above" (background class) is effectively an implicit class. Being implicit, there are no training data provided for it, we end up handling it as a problem of trying to identify OOD through thresholding criteria. There is a key feature here. In OOD determination, there is no discriminative component of the model assigned to the class – the rejection is effectively performed based entirely on low likelihood as computed from the distributions of the known classes, and as a consequence heuristics must be imposed. When we convert "none of the above" to an explicit class, as we have, it converts this to a discriminative modeling problem. The challenge is that, given the vast scope of the "none of the above" class, it is generally infeasible to obtain sufficient training data to model all possibilities. This is a known problem.

This has also been noticed in the ReferFormer and MTTR, where, when we introduce the none-of-the-above as an explicit class through a classification head, it provides limited benefit – the ReferFormer is unable to model it well.

Our cyclic consistency approach provides us a way to capture this class using just a limited number of training samples from this now-explicit class, and we are able to do this because of the specific nature of the R-VOS problem. This, in fact, is a novelty of our approach – we are exploiting the nature of the problem to be able to model this very diverse class effectively using a limited number of training samples. This also clearly shows up in the performance numbers.

B.3 Difference between Our Method and ReferFormer

We summarize the difference between our method and ReferFormer as follows.

- Different from all the existing R-VOS methods, including ReferFormer, using all positive text-video pairs for training, we use both positive and negative pairs, which help the learning of differentiating semantic consensus between different pairs.
- We leverage the relational text-video-text cycle consistency to better correspond the text embedding space to the video embedding space. Positive pairs are constrained with the cycle consistency for better embedding learning, while negative pairs unconstrained with the cycle constraint could be identified.
- We utilize the early-grounding module, which modulates the video feature with the video-aware text embedding. Thus, irrelevant video features are suppressed in an early stage, while ReferFormer only uses dynamic convolution in the final mask decoding stage, easier to involve irrelevant objects, as shown in the results of positive pairs Figure 6.
- Our instance query is composed of both the original sentence embedding and the reconstructed one. Different from ReferFormer that only utilizes original sentence embedding as queries, the reconstructed embedding can encode visual information to facilitate the instance query to decode the objects from visual features.
- Our method achieves superior performance than Referformer. Under the same ResNet-50 backbone, our method achieves 57.3 $\mathcal{J}\&\mathcal{F}$, 94.1 $\mathcal{R}$ and 30 FPS compared to the 55.6 $\mathcal{J}\&\mathcal{F}$ 30.6 $\mathcal{R}$ and 22 FPS of ReferFormer.

B.4 Difference between our Relational Cycle Consistency and Previous Methods (Shi et al., 2020; Chen et al., 2019)

We summarize the difference between our relational cycle consistency and previous related methods (Shi et al., 2020; Chen et al., 2019) as follows.

- We use relational cycle consistency instead of the previous point-wise counterpart, which makes the cycle constraint feasible between two feature spaces that do not have strict bijective mapping, as illustrated in Figure 2 (d). In particular, the mapping from visual objects to textual expressions is not necessarily bijective, as there could be multiple textual descriptions for the same object (about 5 for Ref-Youtube-VOS). Thus, naively adding point-wise consistency may make the feature space collapse. Our ablation study in Table 3 demonstrates the effectiveness of our relational cycle consistency. The point-wise cycle consistency even decreases the accuracy.
- We apply the cycle consistency in the text embedding space instead of the original text expression space (Shi et al., 2020; Chen et al., 2019), which avoids the dataset bias of the pretrained linguistic model from other datasets. Also, we enable the joint optimization of

the primary and dual problem efficiently without decoding text embeddings into expressions, as illustrated in Figure 2 (b).

- We enable a joint optimization of the primary referring segmentation and dual text reconstruction task by introducing a intermediate proxy from early grounding module, thus avoiding redundant two-stage training, to save cost.

C More Implementation Details

We pretrain our model on a combination of three image-level datasets, i.e., Ref-COCO, Ref-COCO+, and Ref-COCOg (Yu et al., 2016). To be compatible with the image-level dataset, we set the window size to 1. We pretrain our model for 12 epochs, which takes about 1-2 days on 8 NVIDIA V100 32G GPUs depending on the backbones. We select the checkpoint with the best results on Ref-COCO val set as our pretrained weight for our main training.

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. $C_v = C_e = C_q = 256$ is utilized. The positional embedding added in the transformers is the standard triangle positional embedding used in (Vaswani et al., 2017). We set the layer number to three for transformers decoders $\mathcal{D}_e$ and $\mathcal{D}_v$. The dynamic filter number K is set to 3. The data point to calculate the relational loss is selected within each batch for simplicity. The text encoder is frozen during the main training.

D Detailed Structure of Mask Decoding

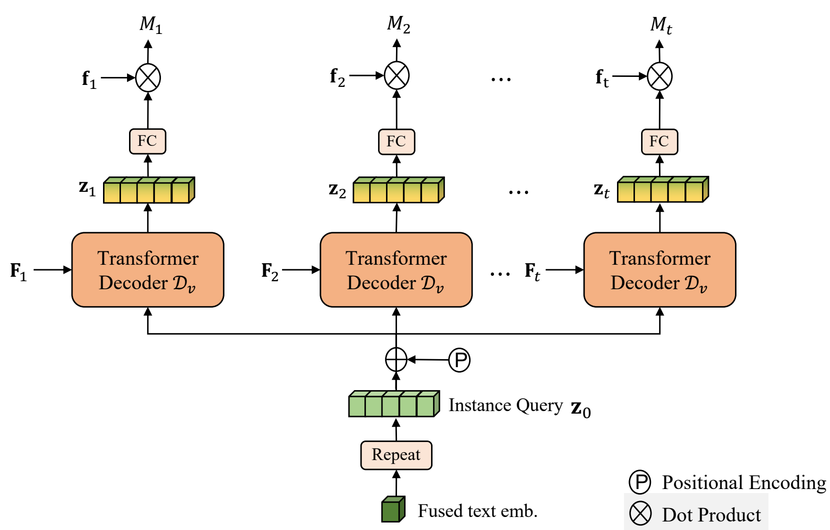

Figure A: **Illustration of mask decoding.**

As is shown in Fig. A, given the fused text embedding, we generate the instance query $\mathbf{z}_0$ by repeating the fused text embedding N times where N is the query number. After that, we generate instance embedding $\{z_t\}_{t=1}^{T}$ for each time step separately using a shared transformer decoder $\mathcal{D}_v$ with encoded memory $\{\mathbf{F}_t\}_{t=1}^{T}$ from visual encoder. The mask prediction M_t for each time step t is derived by a linear combination of $\mathbf{F}_t$ where weights are learned from instance embedding $\mathbf{z}_t$ by two fully connected layers. Note that, as positional embedding is added to the instance query $z_0 \in \mathbb{R}^{C_q \times N}$, each slot in the instance query is different.

Why use N instance queries for only one referred object in the video? Empirical, each slot in the instance query tends to focus on different visual features in the transformer decode $\mathcal{D}_v$ thus the N slots in the instance embedding are highly specialized. Each slot tends to represent an object with some specific properties. For example, slot 1 can always tend to predict an object located in the left of the frame. Slot 2 tends to predict objects belonging to "cat", "dog", etc., categories. By using more than one slot for the instance query, we can generate more specialized and accurate instance embedding, which is vital for mask decoding and confidence score, and box prediction.

E LIMITATIONS

An important challenge for video segmentation is that target object disappearance due to occlusion, which can results in false positives on a per-frame level. In our method, we predict the video-level semantic alignment to handle the false positive in video-level resulted from unpaired text-video pairs. However, since only video-level object expression is available in RVOS task, our method can not address the frame-level false positives resulted from occlusion.

F BROADER IMPACT AND FUTURE WORKS

The false alarm problem in the RVOS task also exists in other referring prediction tasks, e.g., visual grounding (Deng et al., 2021) and referring image segmentation (Ye et al., 2019). We consider our problem formulation that defines the negative and positive vision-language pairs can be extended to other tasks that require multi-modal semantic consensus.

G MORE VISUALIZATION

G.1 VISUALIZATION OF ATTENTIONS IN THE EARLY GROUNDING MODULE

We visualize the cross-attention attentions and $\mathbf{f_{early}}$ in the Early Grounding Module as shown in Figure B.

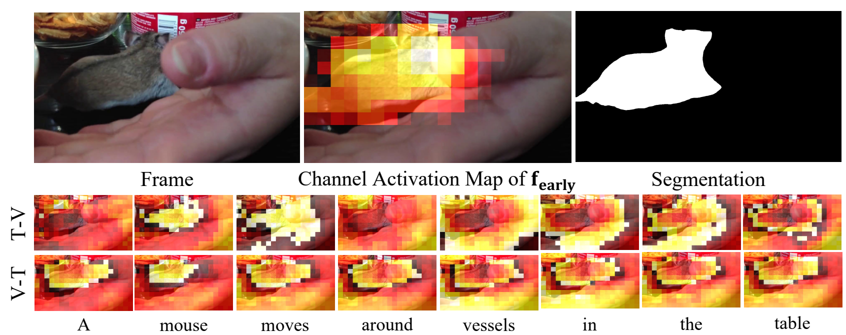

Figure B: **Visualization of cross-attention attentions and $\mathbf{f_{early}}$ in the Early Grounding Module.**

G.2 VISUALIZATION OF SEGMENTATION IN POSITIVE PAIRS

We visualize more segmented masks of positive pairs as shown in Figure C. More visualization for both positive and negative pairs are available in the demo video.

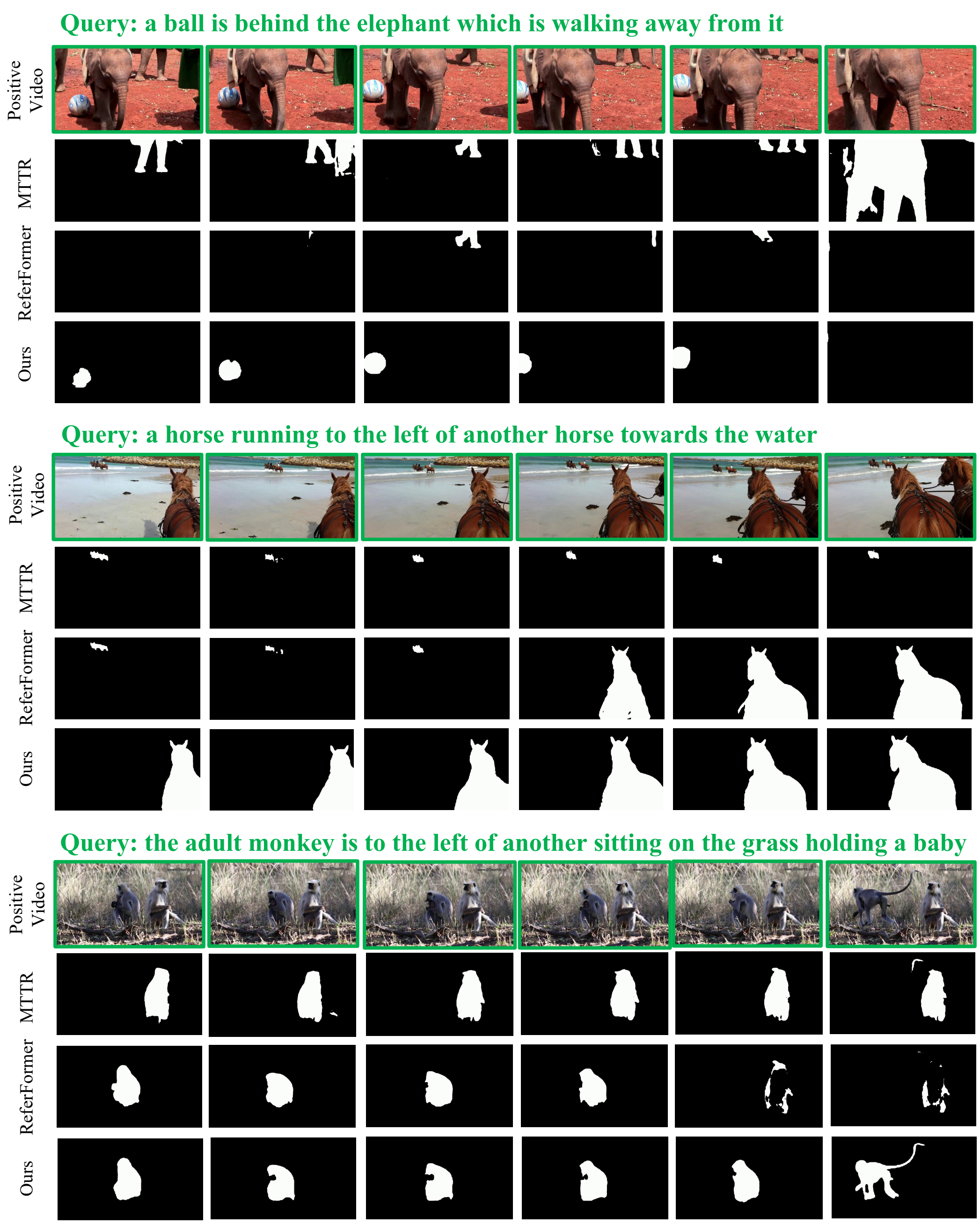

Figure C: **Visualization of Segmentation in Positive Videos.** More visualization are available in the demo video.