Vision-Aware Text Features in Referring Image Segmentation: From Object Understanding to Context Understanding

Hai Nguyen-Truong[1] E-Ro Nguyen[2,3,4*] Tuan-Anh Vu[1†]
Minh-Triet Tran[3,4] Binh-Son Hua[5] Sai-Kit Yeung[1]

[1]The Hong Kong University of Science and Technology [2]Stony Brook University
[3]University of Science, VNU-HCM, Ho Chi Minh City
[4]Vietnam National University, Ho Chi Minh City
[5]Trinity College Dublin

Abstract

Referring image segmentation is a challenging task that involves generating pixel-wise segmentation masks based on natural language descriptions. The complexity of this task increases with the intricacy of the sentences provided. Existing methods have relied mostly on visual features to generate the segmentation masks while treating text features as supporting components. However, this underutilization of text understanding limits the model's capability to fully comprehend the given expressions. In this work, we propose a novel framework that specifically emphasizes object and context comprehension inspired by human cognitive processes through Vision-Aware Text Features. Firstly, we introduce a CLIP Prior module to localize the main object of interest and embed the object heatmap into the query initialization process. Secondly, we propose a combination of two components: Contextual Multimodal Decoder and Meaning Consistency Constraint, to further enhance the coherent and consistent interpretation of language cues with the contextual understanding obtained from the image. Our method achieves significant performance improvements on three benchmark datasets RefCOCO, RefCOCO+ and G-Ref. Project page: https://vatex.hkustvgd.com/.

1. Introduction

Referring image segmentation (RIS) is an emerging new task in computer vision that predicts pixel-wise segmentation of visual objects in images from natural language cues.

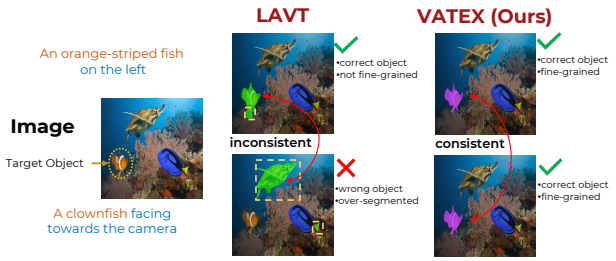


Figure 1. Qualitative comparison between LAVT and Ours. The yellow box indicates the wrong segmentation results. Object understanding and Context understanding are required to tackle the challenge of complex and ambiguous language expression.

Compared to traditional segmentation [5, 16, 50, 52, 55, 57, 61], RIS allows users to select and control the segmentation results via text prompts, which is useful in various applications such as image editing, where users can modify specific parts of an image using simple text commands, and in robotics, where robots need to understand and act on verbal instructions in dynamic environments.

A particular technique to solve RIS is to obtain a robust alignment between language and vision. Performing such an alignment presents significant challenges due to the nature of languages, which are highly ambiguous without given context. Early alignment approaches [13, 31, 41, 63] in RIS either used bottom-up methods, merging vision and language features in early fusion and using an FCN [46] as a decoder to produce object masks, or top-down methods, which first identify objects in image and use the expression as the grounding criterion to select best-matched result.

Recent approaches [11, 36, 62] are based on transformers that learn the interaction between vision-text modalities followed by a standard encoder-decoder process to produce

*Co-first author
†Corresponding author

pixel-level segmentation results. However, existing methods have relied mostly on visual features to generate the segmentation masks while treating text features as supporting components in the fusion module. This insufficient utilization of text understanding hampers these methods' ability to accurately segment target objects for challenging expressions involving rare object vocabulary or contextual relationships between objects. For instance, as illustrated in Figure 1, while LAVT [62] can effectively segment the well-defined "orange-striped fish" and identify specific location information "on the left" in the first expression, it struggles with the expression "clownfish" and incorrectly identifies the turtle that is "facing towards the camera" in the second expression. This inconsistency highlights the limitations of current approaches in understanding complex expressions, especially in handling unseen vocabulary and varying expressions referring to the same object.

Moreover, the human approach to RIS [1, 2, 23] does not involve parsing or understanding complex sentences entirely. Instead, we naturally break down a referring expression into its core components: the object of interest and its description with context information. Initially, the primary focus is on identifying what is the object mentioned in the expression (e.g., the main object of interest). Following this, the search space within the image is narrowed to objects that match the main object's category. The final step involves using the specific characteristics or contextual information described in the expression to pinpoint the target object. Inspired by this, we propose decomposing this task into two processes: object understanding and context understanding. This decomposition allows for a more comprehensive understanding of text features, ultimately enhancing the accuracy and consistency of referring expression segmentation.

Firstly, in terms of *object understanding*, current methods do not utilize the object representation in the query initialization process. ReferFormer [54] generated object queries conditioned on language expressions, while VLT [10] implicitly employed multiple query vectors with different attention weights to generate various interpretations of the language description. However, these variations may lead to confusion and conflict with each other and may not focus on the target object. On the other hand, we propose CLIP Prior to explicitly integrate visual information of the primary object of interest into text cues during the query initialization process. This module transfers the knowledge from pre-trained model CLIP [44] and generates an object-centric visual heatmap to create adaptive, vision-aware queries, enhancing generalization and robustness of object comprehension, even in the challenging case where the expression contains "unseen" category (*e.g.* clownfish).

Secondly, for *context understanding*, we introduce a Contextual Multimodal Decoder (CMD) to further exploit the superior interaction between visual and text modalities, especially the vision-to-language interaction. CMD aims to enhance text features by using contextual information obtained from the visual features and to bring the semantic-aware textual information back to visual features in a hierarchical architecture. While we can use the ground truth mask annotations as a direct learning signal to supervise the language-to-vision features, the opposite interaction is implicitly learned without any learning signal. By observing that there are multiple ways to describe an instance based on the context provided by the image, we propose the Meaning Consistency Constraint (MCC) as a contrastive learning signal to enforce the consistency of vision-aware text features produced from CMD among different expressions referring to the same instance in an image. The vision-to-language interaction can explicitly learn through this extra in-context learning signal, resulting in a *profound, coherent, and contextual understanding* in the feature space.

Our method is evaluated on three widely-used image datasets, RefCOCO, RefCOCO+, and G-Ref, and further extends the results to video datasets, Ref-Youtube-VOS and Ref-DAVIS17. These datasets consist of diverse and challenging text expressions, and our proposed model achieves state-of-the-art performance on all five. Through various ablation studies, we have demonstrated the effectiveness of our model and shown that it can achieve robust referring segmentation even in challenging scenarios.

Our main contributions can be summarized as follows:

- We address the current limitations of existing methods in dealing with complex text expressions and present a novel framework to utilize **V**ision-**A**ware **TEX**t Features (**VATEX**) for a better understanding of text expressions in RIS by decomposing it into Object Understanding and Context Understanding components.
- We introduce a novel CLIP Prior to embed an object-centric visual heatmap in the query initialization process, enhancing object understanding by transferring knowledge from the pre-trained CLIP model.
- We propose Contextual Multimodal Decoder (CMD) followed by a Meaning Consistency Constraint (MCC) as a learning signal for vision-to-language branch to improve context understanding. CMD module enhances the interaction between visual and text modalities, while MCC ensures consistent interpretation of different expressions conditioned in an image.
- Our method achieves superior performance on all splits of the RefCOCO, RefCOCO+, and G-Ref for image datasets and Ref-YouTube-VOS and Ref-DAVIS 2017 for video datasets, surpassing the current state of the art for each dataset, especially in datasets with the more complex expressions.

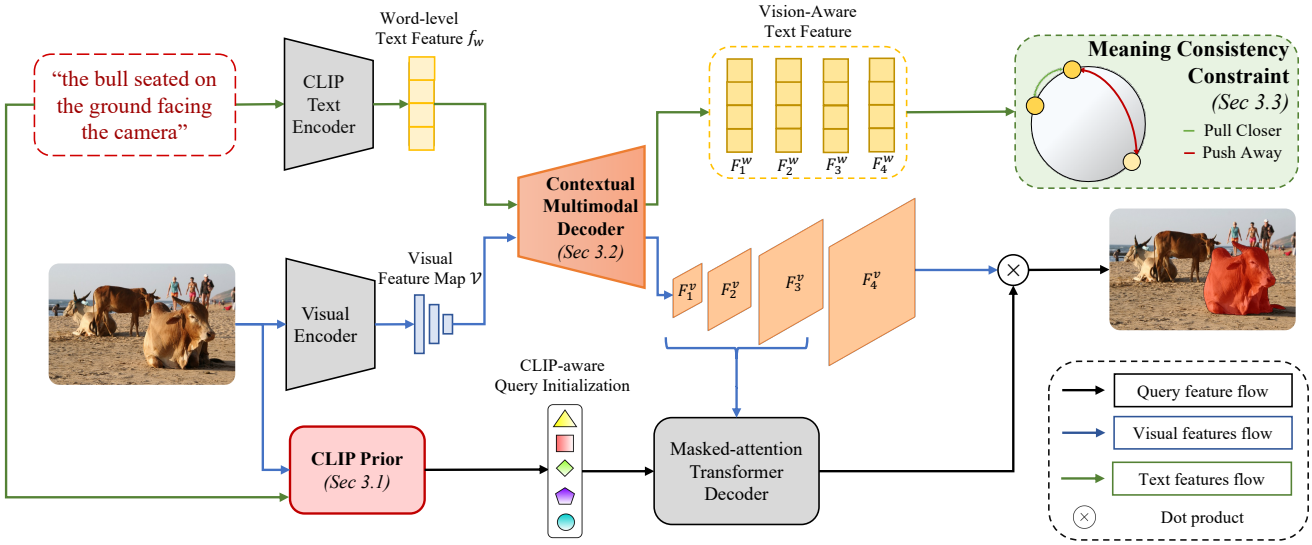


Figure 2. The overall architecture of VATEX, which processes input images and language expressions through two concurrent pathways. Initially, the CLIP Prior module generates object queries, while simultaneously, traditional Visual and Text Encoders create multiscale visual feature maps and word-level text features. These visual and text features are passed into the Contextual Multimodal Decoder to enable multimodal interactions, yielding vision-aware text features and text-enhanced visual features. We then harness vision-aware text features to ensure semantic consistency across varied textual descriptions that reference the same object by employing sentence-level contrastive learning, as described in the Meaning Consistency Constraint section. On the other hand, the text-enhanced visual features and the object queries generated by the CLIP Prior are refined through a Masked-attention Transformer Decoder to produce the final output segmentation masks.

2. Related Work

Referring image segmentation [18] aims to generate pixel-wise segmentation masks for referred objects in images given a natural language expression. Early works [14, 19, 35] proposed to extract visual and linguistic features independently from convolutional and recurrent neural networks, respectively, and then concatenating these features to create multimodal features for decoding final segmentation results. In recent works [3, 10, 12, 30, 51, 58, 62, 64, 66, 68], transformer-based multimodal encoders have been designed to fuse visual and text features, capturing the interaction between vision and language information in the early stage. VG-LAW [47] utilizes language-adaptive weights to dynamically adjust the visual backbone, enabling expression-specific feature extraction for better mask prediction, while LISA [29] learns to generate segmentation masks based on Large Language Models. PolyFormer [36], on the other hand, treats this task as a sequential polygon generation. Another line of works have explored enhancing text understanding in RIS using graph-based methods: CMPC-RefSeg [20] classifies words into entity, attribute, relation, and other categories, building a graph with entities and attributes as nodes and relations as edges, while LSCM-RefSeg [22] constructs fully connected graph, then based on dependency parsing trees to prune unnecessary edges. In contrast, our model does not depend on graph convolutional networks, instead focusing on utilizing the vision-aware text features in text understanding.

Query Initialization. The DETR (DEtection TRansformer) framework [4] has achieved impressive performance in object detection by directly transforming the task of object detection into a set prediction problem. Building upon DETR, several works have focused on improving the query initialization process for better performance. Deformable DETR [67] proposes a deformable transformer architecture to refine object queries, while DAB-DETR [37] directly uses the bounding box coordinate in the image to improve query initialization. In the field of referring segmentation, ReferFormer [54] extracts the word embeddings from the referring expression and treats them as the initial query for the framework. Our CLIP Prior elevates this approach by incorporating a CLIP-generated heatmap, enriching the textual features with spatial context during query initialization. This enriched query leads to improved performance in the subsequent segmentation stages.

Contrastive Learning is pivotal in advancing vision-language tasks [6, 7, 15, 56], enhancing model performance by distinguishing similarities and differences in visual and textual data. CLIP [44] employed a contrastive loss on an extensive image-text dataset. CRIS [53] leveraged text and pixel-level contrastive learning while VLT [11] applied

masked contrastive learning to refine visual features across diverse expressions. Unlike previous approaches [11, 53] that solely focus on improving visual qualities by raw linguistic information, our work utilizes contrastive learning to enhance the comprehension of varied expressions conditioned in a shared image context before using it to enrich the visual features. This ensures the accuracy and stability of mutual interaction between text and visual features, particularly through the comparison of vision-aware expressions related to objects in an image.

3. Proposed Method

Inspired by the human approach to RIS, which involves breaking down a referring expression into its core components: object of interest and contextual description, we propose simplifying the text expression by decomposing it into object and context parts. This decomposition aims to enhance the text understanding, thus improving the accuracy and consistency of referring expression segmentation.

Our framework is constructed by three main components, as demonstrated in Figure 2. First, for object understanding, we propose a CLIP Prior module to generate an object-centric visual heatmap that localizes the object of interest from the text expression, which can be subsequently used to initialize the object queries for the DETR-based method (Section 3.1). Next, we utilize cross-attention modules to interact between visual-text modalities in a hierarchical architecture via our Contextual Multimodal Decoder (Section 3.2) and leverage Meaning Consistency Constraint to harness vision-aware text features generated by CMD (Section 3.3). We further adopt a masked-attention transformer decoder [8] to enhance the object queries through multiscale text-guided visual features. Finally, the enhanced object queries and the visual features from CMD are utilized to output segmentation masks (Section 3.4).

Mathematically, given an input image with the size of $H \times W \times 3$, we can obtain the multiscale visual feature maps $\mathcal{V} = \{V_i\}_{i=1}^{4}, V_i \in \mathbb{R}^{H_i \times W_i \times C_i}$ from the Visual Encoder that captures the visual information in the data, where H_i, W_i, C_i denote the height, width, and the channel dimension of V_i. Given the L-word language expression as input, we use our Text Encoder to encode it into word-level text features $f_w \in \mathbb{R}^{L \times C}$ with C as the channel dimension. Our visual and text features will be further processed as described in the following sections.

3.1. Object Localization with CLIP Prior

For the object part, we first extract the main noun phrase from the expression using spaCy [17] (e.g., the bull) in order to focus only on this main noun phrase. Referring expressions in the RIS task are object-centric, which means that the main noun phrase is the main object of the sentence. We then convert the complex referring expression to a simple template-based sentence before passing it to the CLIP Encoder. In our implementation, we use "A Photo of [Object]" as our template as it is the most common prompt to describe an object [65] in CLIP, where the resulting text feature is represented by $\mathcal{F}_{\text{text}}$. We found that this improves the accuracy of the heatmap in localizing the object of interest. In a separate flow, our input image goes through the CLIP Visual Encoder, resulting in features for multiple image tokens $\mathcal{F}_{\text{CLIP-Image}} \in \mathbb{R}^{\left(\frac{H}{16} \times \frac{W}{16} + 1\right) \times C}$.

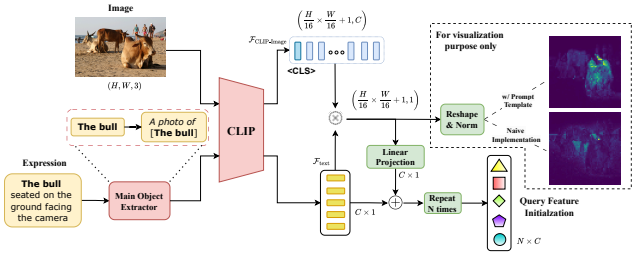


Figure 3. Our CLIP Prior exploits the alignment of CLIP-Image and CLIP-Text embeddings for better query initialization. Best viewed with zoom.

Our visual heatmap for the object of interest can be obtained by calculating the similarity between the visual and text features, then reshaping to image space and going through L2-normalization:

$$\text{Heatmap} = \text{norm}\left(\frac{\mathcal{F}_{\text{CLIP-image}}}{\|\mathcal{F}_{\text{CLIP-image}}\|} \cdot \frac{\mathcal{F}_{\text{text}}}{\|\mathcal{F}_{\text{text}}\|}\right). \quad (1)$$

As normal practice, positional prior from CLIP is embeded to the text features by changing the dimension of the similarity map from $\frac{H}{16} \times \frac{W}{16} + 1$ to C, then repeat it N times together with the text features, then add these two to create initial object queries feature with N queries, each with C-dimension, for embedding the positional prior and text information into the query feature. Unlike previous methods that typically learn the target object representation *implicitly* through multi-modal transformers [10] or convert only *textual information* from natural language into object queries [54], our approach explicitly generates the heatmap and embeds it in the query initialization process. This ensures that the queries contain rich and useful information about both visual and textual aspects, as well as the alignment between these modalities. Such a comprehensive query initialization is essential for effective object understanding, particularly in challenging scenarios involving complex or unseen vocabulary.

While this template approach can efficiently localize regions of interest, it may lead to information loss due to oversimplification (e.g., focusing only on the bull in this scenario). However, its primary function is to narrow down the search space by *localizing a region* containing the object of interest, not necessarily finding out the exact object. To find

the exact object of interest, the full-text expression needs to pass through the CMD and MCC for more comprehensive characteristics and contextual understanding.

3.2. Contextual Multimodal Decoder

Contextual Multimodal Decoder (CMD) is proposed to produce multi-scale text-guided visual feature maps while enhancing contextual information from the image into word-level text features in a hierarchical design, see the architecture figure in the supplementary material. Specifically, our CMD is based on a feature pyramid network architecture [33], which has four levels. Each level transfers the semantic information from visual features to text features and then uses these vision-aware text features to update the visual features afterward via cross-modal attention.

In the i-th level of CMD, given the input visual features V_i and text features F_{i-1}^w, the multi-modal interactions are performed in two steps. First, a cross-attention that takes text features F_{i-1}^w as query and visual features V_i as key and value is used to model the relationship of the text and visual information. Then it forms the vision-aware text features by associating them with current text features:

$$F_i^w = \text{MHA}(F_{i-1}^w, V_i, V_i) \cdot F_{i-1}^w, \tag{2}$$

where $\text{MHA}(q, k, v)$ is the multi-head cross-attention module with query q, key k, value v.

F_i^w is then treated as the key and value and V_i is treated as the query in another multi-head cross-attention module to reinforce the alignment between the visual and text modalities and generate features V_i'. Consequently, V_i' is fused with the text-guided visual feature F_{i-1}^v from the previous level $i-1$ followed by a Conv2d layer to obtain the text-guided visual feature F_i^v. Mathematically, the whole process is as follows:

$$V_i' = \text{MHA}(V_i, F_i^w, F_i^w) \cdot V_i, \tag{3}$$
$$F_i^v = \text{Conv2d}(V_i' + \text{Ups}(F_{i-1}^v)), \tag{4}$$

where $\text{Conv2d}()$ is the 2D convolutional layer, and $\text{Ups}()$ denotes upsampling F_{i-1}^v to the size of V_i'. For the first level with $i = 1$, we skip F_0^v and let $F_0^w = f_w$, where f_w is the word-level linguistic features extracted by the Text Encoder.

Previous methods have developed various bidirectional multimodal fusion mechanisms, including word-pixel alignment in encoder stage [58, 64] and region-language interactions [34]. Compared to these approaches, our novelty is the combination of CMD and MCC, where MCC serves as an in-context learning signal to enrich the vision-aware text features and further enhance the vision-language interactions within the hierarchical architecture of CMD.

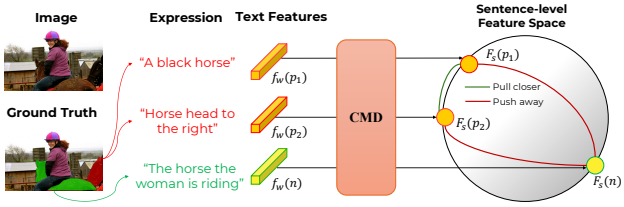


Figure 4. Illustration of Meaning Consistency Constraint. Vision-aware text embeddings of different expressions are passed through a contrastive learning module in sentence-level feature space. Embeddings referring to the same object are pulled closer while pushing others far away. Best view in color.

3.3. Meaning Consistency Constraint

Each object in an image can be described by various text expressions. Although the linguistic meaning of these descriptions may be different, they should convey the same semantic meaning when referencing that image (see two red expressions in Figure 4). According to this perspective, it's crucial for CMD to gradually comprehend contextual cues from visual features into textual representations and ensure consistent identification of target objects, where expressions referring to the same object yield identical representations. However, previous studies have often overlooked the relationship between expressions that pertain to the same instance.

To delve deeper into this relationship and provide the explicit in-context learning signal for vision-aware text features within CMD, we propose Meaning Consistency Constraint (MCC), a sentence-level contrastive learning approach. MCC aims to learn meaningful and discriminative representations for different expressions while consistently pulling sentences referring to the same object close to each other.

Unlike previous contrastive learning-based approaches [11, 53], we focus on linguistic features that are enriched and conditioned by visual information. This can encourage CMD module to gradually learn how to produce richer text features and lead to the improvement of visual features in context understanding due to the bidirectional attention mechanism and hierarchical design of CMD module.

Our contrastive learning pipeline is illustrated in Figure 4. During training, we construct a triplet of text expressions for each image. Each triplet comprises two sentences that refer to the same object (positive samples), along with a third sentence describing a different object (negative sample). We denote the positive samples by p_1, p_2 and the negative sample by n, respectively. With each sample x, we derive the sentence-level feature by averaging the vision-aware word-level textual features:

$$F_s(x) = \text{Avg}(F_4^w(x), \text{dim} = 0), \tag{5}$$

Table 1. Quantitative results of referring image segmentation on Ref-COCO, Ref-COCO+, G-Ref datasets on mIoU metric.

Method	Backbone		RefCOCO			RefCOCO+			G-Ref	
	Visual	Textual	val	testA	testB	val	testA	testB	val	test
CRIS [53]	ResNet-101	CLIP	70.47	73.18	66.10	62.27	68.08	53.68	59.87	60.36
CM-MaskSD [49]	CLIP-ViT-B	CLIP	72.18	75.21	67.91	64.47	69.29	56.55	62.67	62.69
VLT [11]	Swin-B	BERT	72.96	75.96	69.60	63.53	68.43	56.92	63.49	66.22
LAVT [62]	Swin-B	BERT	74.46	76.89	70.94	65.81	70.97	59.23	63.34	63.62
LISA-7B [29]	ViT-H SAM	Vicuna-7B	74.10	76.50	71.10	65.10	67.40	56.50	66.40	68.50
VG-LAW [47]	ViT-B	BERT	75.05	77.36	71.69	66.61	70.30	58.14	65.36	65.13
VATEX (Ours)	Swin-B	CLIP	**78.16**	**79.64**	**75.64**	**70.02**	**74.41**	**62.52**	**69.73**	**70.58**

Table 2. Precision analysis at different threshold value comparison between VATEX and recent SOTA methods.

Methods	Pr@0.5	Pr@0.7	Pr@0.9	mIoU
LAVT [62]	84.46	75.28	34.30	74.46
ReLA [34]	85.92	77.71	34.99	75.61
CG-Former [48]	87.23	78.69	38.77	76.93
VATEX (Ours)	**88.12**	**82.54**	**45.11**	**78.17**

Table 3. Quantitative results on video datasets.

Methods	Ref-YT-VOS			Ref-DAVIS17		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
ReferFormer [54]	62.9	61.3	67.5	61.1	58.1	64.1
VLT [11]	63.8	61.9	65.6	61.6	58.9	64.3
VATEX (Ours)	**65.4**	**63.3**	**67.5**	**65.4**	**62.3**	**68.5**

where $F_4^w \in R^{L\times C}, F_s \in R^C$. We adopt the InfoNCE loss [43] to ensure that linguistic features referring to the same object converge, while features of different objects diverge:

$$\mathcal{L}_{mcc} = -\log\left(\frac{\text{sim}(p_1,p_2)}{\text{sim}(p_1,n)+\text{sim}(p_2,n)}\right), \quad (6)$$

where $\text{sim}(p,n) = \exp(F_s(p)\cdot F_s(n))$ to calculate the exponential for cosine similarity of sentence-level obtained from the text expressions.

3.4. Network Training

Prediction Heads. We adopt the masked-attention transformer decoder [8] to update our initial query feature f_o by using the multi-scale text-guided visual features $\{F_i^v\}_{i=1}^3$ to obtain the final object queries $F_o \in \mathbb{R}^{N\times C}$. The final object queries will directly predict the probability of the target object $\hat{p} \in \mathbb{R}^N$. The high-resolution segmentation mask $\hat{s} \in R^{\frac{H}{4}\times\frac{W}{4}\times N}$ is produced by associating between object queries F_o with the last fine-grained text-guided visual features $F_4^v \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C}$, which can be formulated as:

$$\hat{s} = \text{Sigmoid}(F_4^v \cdot F_o^\top). \quad (7)$$

Instance Matching. The prediction set output from prediction heads is represented by $\hat{y} = \{\hat{y}_i\}_{i=1}^N$, where $\hat{y}_i = \{\hat{p}_i, \hat{s}_i\}$. Since a text expression refers to only a specific object, we denote the ground truth object as $y = \{p_{gt}=1, s_{gt}\}$. The best prediction $\hat{y}_\delta$ can be found by a Hungarian algorithm [28] by minimizing the matching cost in terms of probability and segmentation mask [8, 9].

Training. Our prediction $\hat{y}_\delta$ is supervised by three losses. Firstly, the class loss $\mathcal{L}_{cls}$ is binary cross entropy (BCE) loss to supervise the probability of the referred object. Secondly, the mask loss $\mathcal{L}_{mask}$ is a combination of dice loss and BCE. Finally, our sentence-level contrastive loss $\mathcal{L}_{mcc}$ is used to enforce our Meaning Consistency Constraint. The total loss can be formulated as follows:

$$\mathcal{L}_{total} = \gamma_{cls}\mathcal{L}_{cls} + \gamma_{mask}\mathcal{L}_{mask} + \gamma_{mcc}\mathcal{L}_{mcc}, \quad (8)$$

where $\gamma_{cls}, \gamma_{mask}, \gamma_{mcc}$ are the scalar coefficients.

Inference. In inference, our method aligns with the standard practice of using a single image or video with one text expression, and MCC only requires sampling positive and negative expressions in the training phase. During inference, the query with the highest probability score is selected as the target object for the final output.

4. Experimental Results

4.1. Experiment Setup

We evaluate the performance of our model on three image datasets: RefCOCO [24], RefCOCO+ [24], G-Ref [42] and further evaluate the performance of our model on two video datasets: Ref-Youtube-VOS [45] and Ref-DAVIS17 [25]. For evaluation metrics, we follow previous work to use mean IoU (mIoU) and Precision at different thresholds (Pr@X) for image and $\mathcal{J}\&\mathcal{F}$ for video datasets.

During training, we freeze the CLIP model. Images are resized to a short side of 480. We set the coefficients for the losses as $\gamma_{cls}=2, \gamma_{mask}=5$, and $\gamma_{mcc}=2$, with the feature dimension C set to 256. We train the network for $100,000$ iterations on the RefCOCO(+/g) datasets with an initial learning rate of 10^{-4} and is reduced by a factor of 0.1 at the $2/3$ last iteration. For the Ref-Youtube-VOS dataset, we initialize the pre-trained weight from RefCOCO(+/g) and train the network for $100,000$ iterations. Regarding the Ref-DAVIS17 dataset, we directly use the weight obtained from the Ref-Youtube-VOS dataset for inference. The training process uses 2 NVIDIA RTX 3090 GPUs with a batch size of 24. For detailed information on each dataset and implementation, please see the supplementary material.

Figure 5. Results on RefCOCO(+/g) datasets. We compare our results with CRIS and LAVT. Our method excels at segmenting objects in complex scenarios, such as distinguishing between similar objects and localizing specific instances within a scene. The last two columns of the results show failure cases. Best viewed in color.

Table 4. Ablation Study on the val set of RefCOCO (mIoU) and Ref-YouTube-VOS ($\mathcal{J}\&\mathcal{F}$).

CLIP Prior	CMD	MCC	RefCOCO	Ref-YT-VOS
-	-	-	70.42	59.8
✓	-	-	71.95 +1.53	61.5 +1.7
-	✓	-	73.18 +2.76	61.9 +2.1
-	✓	✓	75.43 +5.01	63.6 +3.8
✓	✓	✓	**78.16** +7.74	**65.4** +5.6

Table 5. Ablation on different query initialization methods in CLIP Prior.

Query Initialization method	RefCOCO
Only text features [54]	75.43 -2.73
CLIP Prior with Prompt "A Photo of [Object]"	78.16
CLIP Prior mean's performance over 80 ImageNet prompts	78.25 +0.09
CLIP Prior w.o main object extractor	74.34 -3.82

Table 6. Ablation on different bidirectional multimodal fusion modules.

Bidirectional fusion	MCC	RefCOCO
CMD (Ours)	✓	**78.16**
ETRIS [58]	✓	77.22 -0.94
CoupAlign [64]	✓	77.01 -1.15
CMD	✗	75.12 -3.04
ETRIS [58]	✗	74.12 -4.04
CoupAlign [64]	✗	73.97 -4.19

4.2. Main Results

Referring Image Segmentation. As illustrated in Table 1, our method outperforms the state-of-the-art methods by a large margin in all splits of different datasets in the standard setting. Notably, our method surpasses the recent CGFormer and VG-LAW on the validation splits of all three benchmarks, achieving mIoU gains of 1.23% and 3.11% on RefCOCO, 1.46% and 3.31% on RefCOCO+, and 2.16% and 4.37% on G-Ref. The more complex the expressions, the greater the performance gains achieved by VATEX. Compared to LISA [29], a large pre-trained vision and text encoder, VATEX consistently outperforms it by 3-5% across all datasets. Furthermore, Table 2 demonstrates the superior performance of VATEX over LAVT, ReLA, and CG-Former on average precision metrics, particularly at the Pr@0.7 and Pr@0.9, illustrating our ability to generate high-quality and complete segmentation masks.

Referring Video Segmentation. Our model can be extended to video datasets with minor adaptations to handle temporal information. As shown in Table 3, VATEX outperforms current SOTA methods VLT and ReferFormer on the same Video-Swin-B backbone by 1.6 and 3.8 $\mathcal{J}\&\mathcal{F}$ on Ref-Youtube-VOS and Ref-DAVIS17 datasets, respectively.

Qualitative Analysis. We provide the visualization of our results in Figure 5. Our method can successfully segment objects in complex scenarios, such as the presence of multiple similar objects. For example, in the first column, we can localize the guy who sits in the chair instead of the man standing on the tennis court. In the second sample, VATEX can distinguish the sushi plate among several food dishes

that LAVT cannot. In the fourth column, our model can not only identify the correct umbrella belonging to the guy in the light jacket but also segment a part of the shaft of the umbrella that the ground truth does not provide. With the expression "bear child is hugging" in the sixth image, LAVT can only segment the bear's head, and CRIS over-segment to the child and under-segment the bear, but VATEX can output the target bear concisely. However, our method fails to segment objects that need to be counted and selected by their order or be described indirectly through another object, as we have not leveraged counting information and object interaction in our model. Another point worth mentioning is the differences in architecture design between VATEX and LAVT. VATEX focuses on instance-based segmentation, while LAVT focuses on pixel-based segmentation. Consequently, VATEX produces smoother and more complete segmentation masks.

4.3. More Analysis

Ablation Study. We conduct an ablation study on the validation sets of RefCOCO with Swin-B backbone and Ref-Youtube-VOS validation set with Video-Swin-B backbone to examine the impact of each proposed component in our model. The baseline follows the architecture of ReferFormer [54] by using only languages as the initial query (removing CLIP Prior), while only using text-guided vision features and ignoring the vision-aware text features (removing CMD and MCC). As shown in Table 4, the combination of CLIP Prior, CMD, and MCC modules results in the best performance, showcasing a remarkable performance increase of up to 7.74% in mIoU on RefCOCO and 5.6% in $\mathcal{J}\&\mathcal{F}$ on Ref-Youtube-VOS. This outcome unequivocally attests to the remarkable effectiveness of our approach and underscores its significant impact. The full ablation study is shown in the supplementary material.

Study on different query initialization methods of CLIP Prior. As described in Section 3.1, our CLIP Prior relies on a template to convert the main noun phrase into a simple sentence suitable for CLIP. As shown in Table 5, the baseline follows the query initialization from [54], which uses only text features and achieves 75.43% mIoU. We investigate the effects of using different templates and how these affect the final performance. By using the template "A Photo of [Object]", there is a notable improvement of 2.73% in mIoU. We conducted an additional experiment where we aggregated the text embeddings from 80 ImageNet prompts, which has a very minor performance improvement. This demonstrates that the template "A Photo of [Object]" remains a practical choice. We also conduct the experiment that using the whole sentence (instead of the main noun phrase) leads to a significant deterioration in performance at 3.82%. In this case, CLIP introduces noisy activation on various objects based on their discriminative characteristics within the complex sentence, which is harmful to the model.

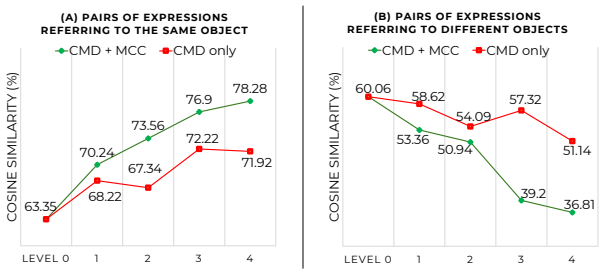


Figure 6. Cosine similarities between sentence-level text features at each CMD level.

Different bidirectional multimodal fusion modules. We ran an ablation study to quantify the performance of CMD with other bidirectional multimodal fusion mechanisms by substituting CMD with alternative modules in ETRIS [58] and CoupAlign [64]. As shown in Table 6, CMD with MCC performs the best, achieving a mIoU of 78.16%, which is higher than both ETRIS [58] and CoupAlign [64] with 0.94% and 1.15%, respectively. Disabling MCC results in a notable drop in performance across all settings, with CMD seeing a decrease of 3.04%, while ETRIS and CoupAlign experience decreases of 4.04% and 4.19%, respectively. This highlights the importance of MCC in improving bidirectional fusion performance, with CMD consistently outperforming alternative methods under the same condition.

Effect of MCC on Vision-aware Text Features. To quantify the impact of the MCC on Vision-aware Text Features, we calculate the similarity between sentence-level text features at each layer of CMD, with and without MCC, using the G-Ref dataset. We chose G-Ref because it contains longer, more diverse, and complex context information about the objects expressions, making it ideal for studying the impact of MCC on Vision-aware Text Features. The average similarity results for all pairs of expressions referring to the same or different objects are depicted in Figure 6. Here, level 0 represents the initial features, derived directly from the text encoder, while level 4 indicates the final vision-aware text features of CMD.

In the context of expressions referring to the same object, the initial similarity of text features stands at 63.35%. By utilizing the MCC, the average similarity gradually increases from level 1 to level 4, reaching 78.28% at the final vision-aware text feature, while without MCC, the similarity score fluctuates between levels 1 and 3, achieving only $\sim$ 8% performance gain at the end. This illustrates the effectiveness of MCC in guiding the vision-aware text features toward a semantically rich feature space, where the features of two sentences referring to the same object are closer in that feature space.

Regarding different objects, the pairwise similarity of expressions referring to different objects progressively decreases through the four levels of CMD, from 60.06% to 36.81%. In contrast, using CMD without MCC results in an unstable feature space. This illustrates the effectiveness of MCC in guiding vision-aware text features toward a more robust feature space, where the features of two sentences referring to different objects can be similar at the beginning but are pushed apart in the final vision-aware text feature.

These findings underscore the pivotal role of MCC in bolstering multimodal comprehension provided by CMD. Specifically, we reveal how the in-context learning signal in MCC effectively improves the representation of vision-aware text features.

Others. We reported the Universality of VATEX, the Runtime Analysis of VATEX, and the effect of MCC on object segmentation in Supplementary Material.

5. Limitations

Our method is not without limitations. Particularly, our method does not exploit positional relationship between different objects as well as the alignment between actions and expressions referring to them (see Figure 5 the last two columns). Consequently, situations involving counting ("third from left"), indirect descriptions ("kid next to girl in pink pants"), or actions ("a woman walking to the left") might lead to inaccurate predictions. Another line of future work is making RIS work on general scenarios (*e.g.* segment all the red-colored objects, segment all the text in the image) or more fine-grained segmentation (*e.g.* segment the eye of the owl). Dealing with intra-frame object relationships and inter-frame information for video is vital for future work. It is also of great interest to investigate vision-aware text features with the VLMs and to lift this task to the 3D domain.

6. Conclusion

This paper introduces VATEX, a novel framework that examines how vision-aware text features can enhance the performance of RIS by emphasizing on object and context comprehension. First, we propose integrating visual cues into text features during the query initialization process in CLIP Prior for object understanding. First, we propose integrating visual cues into text features during the query initialization process via the CLIP Prior module for object understanding. Second, we exploit the mutual interaction between visual and text modalities through the Contextual Multimodal Decoder (CMD) module and provide an explicit in-context learning signal for the vision-to-language direction using the Meaning Consistency Constraint (MCC). As a result, our proposed method consistently achieves new state-of-the-art results on three benchmark datasets: RefCOCO, RefCOCO+, and G-Ref.

Acknowledgment

This research was supported by the internal grant from HKUST (R9429). Binh-Son Hua is supported by the Science Foundation Ireland under the SFI Frontiers for the Future Programme (22/FFP-P/11522).

References

[1] Moshe Bar. A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of cognitive neuroscience*, 15:600–9, 06 2003. 2

[2] Elan Barenholtz. Quantifying the role of context in visual object recognition. *Visual Cognition*, 22(1):30–56, 2014. 2

[3] Adam Botach, Evgenii Zheltonozhskii, Chaim Baskin, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4985–4995, 2022. 3

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

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020. 3

[7] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3

[8] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 4, 6, 13

[9] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021. 6

[10] 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, 3, 4

[11] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vlt: Vision-language transformer and query generation for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 3, 4, 5, 6, 18

[12] 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. 3

[13] Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. Encoder fusion network with co-attention embedding for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15506–15515, 2021. 1

[14] Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. Encoder Fusion Network with Co-Attention Embedding for Referring Image Segmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15501–15510, June 2021. ISSN: 2575-7075. 3

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 06 2020. 3

[16] Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. Vita: Video instance segmentation via object token association. *arXiv preprint arXiv:2206.04403*, 2022. 1

[17] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017. 4, 14

[18] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 108–124. Springer, 2016. 3

[19] Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. Referring Image Segmentation via Cross-Modal Progressive Comprehension. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10485–10494, June 2020. ISSN: 2575-7075. 3

[20] Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. Referring image segmentation via cross-modal progressive comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[21] Ziling Huang and Shin'ichi Satoh. Referring image segmentation via joint mask contextual embedding learning and progressive alignment network. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7753–7762, Singapore, Dec. 2023. Association for Computational Linguistics. 15

[22] Tianrui Hui, Si Liu, Shaofei Huang, Guanbin Li, Sansi Yu, Faxi Zhang, and Jizhong Han. Linguistic structure guided context modeling for referring image segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 59–75. Springer, 2020. 3

[23] Glyn W. Humphreys, Price Cj, and Riddoch Mj. From objects to names: A cognitive neuroscience approach. *Psychological Research*, 62:118–130, 1999. 2

[24] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. 6, 13

[25] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14*, pages 123–141. Springer, 2019. 6, 13

[26] Seoyeon Kim, Minguk Kang, and Jaesik Park. Risclip: Referring image segmentation framework using clip. *arXiv preprint arXiv:2306.08498*, 2023. 18

[27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 16, 17

[28] Harold W. Kuhn. The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, March 1955. 6

[29] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9579–9589, June 2024. 3, 6, 7

[30] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in neural information processing systems*, 34:19652–19664, 2021. 3, 15, 17

[31] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2018. 1

[32] Xiang Li, Jinglu Wang, Xiaohao Xu, Xiao Li, Bhiksha Raj, and Yan Lu. Towards robust referring video object segmentation with cyclic relational consensus, 2023. 18

[33] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection, 2017. 5

[34] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23592–23601, 2023. 5, 6, 15

[35] Daqing Liu, Hanwang Zhang, Zheng-Jun Zha, and Feng Wu. Learning to Assemble Neural Module Tree Networks for Visual Grounding. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4672–4681, Oct. 2019. ISSN: 2380-7504. 3

[36] Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 18653–18663, 2023. 1, 3, 15, 17

[37] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022. 3

[38] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 16, 17

[39] 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. 16

[40] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 13

[41] Gen Luo, Yiyi Zhou, Rongrong Ji, Xiaoshuai Sun, Jinsong Su, Chia-Wen Lin, and Qi Tian. Cascade grouped attention network for referring expression segmentation. *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. 1

[42] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–20, 2016. 6, 13

[43] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 6

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3

[45] 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. 6, 13

[46] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, 2017. 1

[47] Wei Su, Peihan Miao, Huanzhang Dou, Gaoang Wang, Liang Qiao, Zheyang Li, and Xi Li. Language adaptive weight generation for multi-task visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10857–10866, 2023. 3, 6, 18

[48] Jiajin Tang, Ge Zheng, Cheng Shi, and Sibei Yang. Contrastive grouping with transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23570–23580, 2023. 6

[49] Wenxuan Wang, Jing Liu, Xingjian He, Yisi Zhang, Chen Chen, Jiachen Shen, Yan Zhang, and Jiangyun Li. Cmmasksd: Cross-modality masked self-distillation for referring image segmentation. *arXiv preprint arXiv:2305.11481*, 2023. 6, 18

[50] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021. 1

[51] Xudong Wang, Shufan Li, Konstantinos Kallidromitis, Yusuke Kato, Kazuki Kozuka, and Trevor Darrell. Hierarchical open-vocabulary universal image segmentation. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 15, 17, 19

[52] 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. 1

[53] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022. 3, 4, 5, 6, 15, 18

[54] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4984, 2022. 2, 3, 4, 6, 7, 8, 13, 18, 20

[55] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pages 588–605. Springer, 2022. 1

[56] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 3

[57] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 1

[58] Zunnan Xu, Zhihong Chen, Yong Zhang, Yibing Song, Xiang Wan, and Guanbin Li. Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17503–17512, 2023. 3, 5, 7, 8

[59] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15325–15336, 2023. 15, 17

[60] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes

extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023. 14

[61] Shusheng Yang, Xinggang Wang, Yu Li, Yuxin Fang, Jiemin Fang, Wenyu Liu, Xun Zhao, and Ying Shan. Temporally efficient vision transformer for video instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2885–2895, 2022. 1

[62] 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. 1, 2, 3, 6, 15, 18, 20

[63] 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. 1

[64] Zicheng Zhang, Yi Zhu, Jianzhuang Liu, Xiaodan Liang, and Wei Ke. Coupalign: Coupling word-pixel with sentence-mask alignments for referring image segmentation. In *NeurIPS*, 2022. 3, 5, 7, 8

[65] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 2022. 4

[66] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision*, pages 598–615. Springer, 2022. 3, 15, 17

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

[68] Xueyan Zou*, Zi-Yi Dou*, Jianwei Yang*, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee*, and Jianfeng Gao*. Generalized decoding for pixel, image and language. 2022. 3, 16, 17

[69] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024. 16, 17

Vision-Aware Text Features in Referring Image Segmentation: From Object Understanding to Context Understanding

— Supplementary Material —

Our supplementary has 5 sections. Section 7 shows additional information about datasets and training procedure. Section 8 explains how spaCy is used to extract main noun phrases from sentences and also explains how potential LLMs can be used to create diverse object descriptions to improve dataset annotations. Section 9 contains the additional experiments on RefCOCO(+/g), Ref-Youtube-VOS and Ref-DAVIS17. This section also illustrates and analyzes the performance of CLIP Prior, CMD and MCC in different situations as well as the runtime and.

7. Additional Implementation Details

7.1. Datasets

Image datasets. RefCOCO and RefCOCO+ [24] are two of the largest image datasets used for referring image segmentation. They contain 142,209 and 141,564 language expressions describing objects in images. RefCOCO+ is considered to be more challenging than RefCOCO, as it focuses on purely appearance-based descriptions. G-Ref [42], or RefCOCOg, is another well-known dataset with 85,474 language expressions with more than 26,000 images. The language used in G-Ref is more complex and casual, with longer sentence lengths on average.

Video datasets. Ref-YouTube-VOS [45] and Ref-DAVIS17 [25] are well-known datasets for referring video object segmentation. Ref-YouTube-VOS contains 3978 video sequences with approximately 15000 referring expressions, while Ref-DAVIS17 consists of 90 high-quality video sequences. These datasets are used to evaluate the performance of algorithms that aim to identify a specific object within a video sequence based on natural language expressions.

7.2. Metrics

In our work, we use mIoU and Precision@X to evaluate our method for image datasets, while $\mathcal{J}\&\mathcal{F}$ are used as evaluation metrics for video datasets. mIoU stands for mean Intersection over Union, which measures the average overlapping between the predicted segmentation masks and the ground truth annotations. Precision@X, on the other hand, measures the success rate of the referring process at a specific IoU threshold, and it focuses on the referring capability of the method.

In addition, region similarity $\mathcal{J}$ and contour accuracy $\mathcal{F}$, and their average $\mathcal{J}\&\mathcal{F}$ are commonly used evaluation metrics for video object segmentation (VOS) datasets. The $\mathcal{J}$ is similar to the IoU score, while the $\mathcal{F}$ score is the boundary similarity measure between the boundary of the prediction and the ground truth. These two metrics together measure the performance of the predicted object mask over the entire video sequence. Higher $\mathcal{J}\&\mathcal{F}$ score indicates better RVOS performance.

Furthermore, to quantify the ability to consistently segment various expressions for the same object and further validate the effectiveness of our proposed Meaning Consistency Constraint, we leverage an Object-centric Intersection over Union (Oc-IoU) score, which calculates the overlap and union area between ground truth and all segmentation predictions of the same object. Specifically, consider the i-th object with K_i expressions referring to that object and the corresponding ground truth mask GT_i. Let P_i^j be the model's prediction for the j-th expression of the i-th object, where $j = \overline{1..\ K_i}$. The Object-centric IoU can be formulated as follows:

$$\text{Oc-IoU}(\mathrm{GT}_i, \mathrm{P}_i) = \frac{\mathrm{GT}_i \cap \mathrm{P}_i^1 \cap ... \cap \mathrm{P}_i^{\mathrm{K}_i}}{\mathrm{GT}_i \cup \mathrm{P}_i^1 \cup ... \cup \mathrm{P}_i^{K_i}}, \tag{9}$$

$$\text{Oc-IoU}_{\text{total}} = \frac{1}{N} \sum_{i=1}^{N} \text{Oc-IoU}(\mathrm{GT}_i, \mathrm{P}_i), \tag{10}$$

where N is the total number of objects/instances in the datasets.

7.3. Training Details

Our model is optimized using AdamW [40] optimizer with the initial learning rate of 10^{-5} for the visual encoder and 10^{-4} for the rest. Our model comprises a total of nine Masked-Attention Transformer Decoder layers followed [8]. We set the number of queries to 5 [54]. For the setting of training from classification weight from Imagenet on Ref-Youtube-VOS dataset, we train the model for $200,000$ iteration with the learning drop at $140,000$-th iteration. On Ref-DAVIS17 [25], we directly report the results using the model trained on Ref-YouTube-VOS without fine-tuning. In terms of coefficients in loss function, $\gamma_{cls} = 2$ and $\gamma_{mask} = 5$ are followed from Mask2Former. To maintain balance, we then choose $\gamma_{mcc} = 2$. We want to prioritize the mask loss with the highest weight because the IoU is the primary metric.

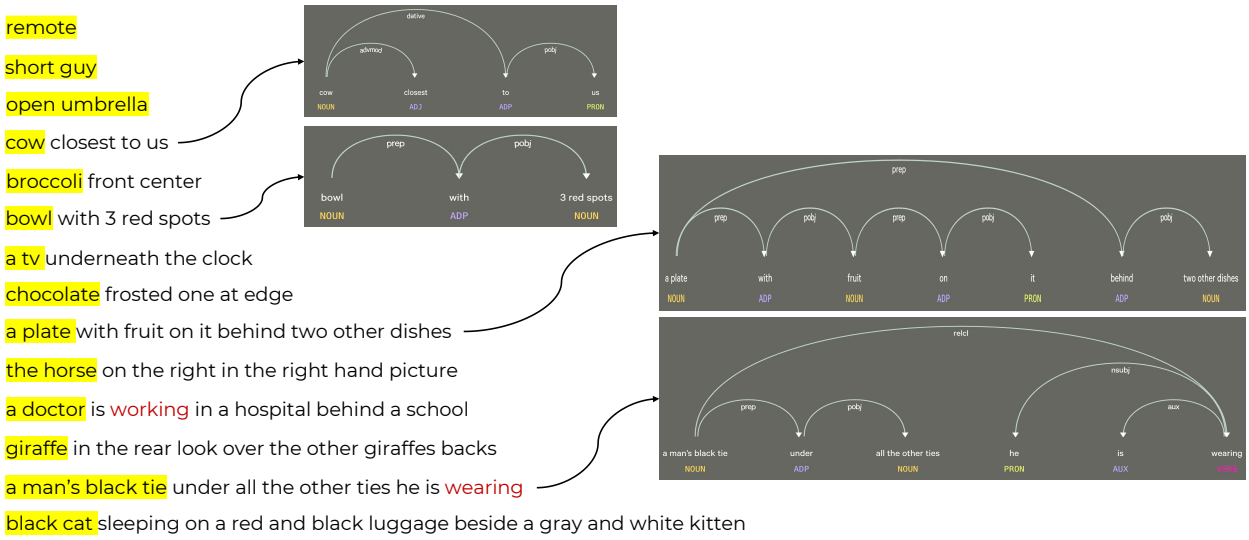


Figure 7. Examples of our main object extractor output. Given the expression, our algorithm will output the main noun phrase in the sentence. Typically, the root word of the sentence is a noun phrase, which we directly output as the main noun phrase. However, if the root word is not a noun phrase (*e.g.* working, wearing in the image), we instead focus on identifying its child noun. Additionally, we illustrate the dependency parsing tree for some representative sentences on the right.

8. Additional Details of VATEX

8.1. Main Object Extractor

We use spaCy [17] to implement our main object extractor, leveraging its optimized, fast, and effective dependency parsing capabilities. First, spaCy extracts the root word of the sentence, also known as the head word, which has no dependency on other words (i.e., it has no parent word in the dependency tree). If this root word is a noun phrase, we directly output it as the main noun phrase of the sentence. If the root word is not a noun (e.g., a verb), we focus on its child noun to ensure it centers on the described object. Figure 7 shows some examples on the datasets and shows the output of our algorithm as well as the dependency parsing tree of some representative cases.

To handle complex sentence structures that lack a directly related noun phrase, we have implemented a rollback mechanism (in L27 of vatex/utils/noun_phrase.py) that returns the whole sentence, preventing information loss and mitigating potential errors from inaccurate main noun phrase extraction. As shown in Table 7, this rollback mechanism helps avoid poorly extracted nouns that could potentially cause incorrect segmentation masks.

8.2. Enhancing Expression Diversity in Referring Image Segmentation Datasets through Prompting Techniques

Our method's utilization of diverse referring expressions for each object aligns with established best practices in text-image dataset annotation. This approach is widely accepted and implemented across several benchmark datasets. In scenarios where multiple expressions per object are unavailable, we have the flexibility to employ Large Language Models (LLMs) for enhancing expression diversity. This can be achieved either by augmenting existing expressions or generating new ones based on object masks, a technique successfully employed by datasets like RIS-CQ. Furthermore, we demonstrate a practical application of this approach through a sample that showcases how we can prompt ChatGPT to generate relevant expressions in Figure 8. This generation is based on factors like an object's position in the image, its relative position to other objects or people, and distinguishing attributes such as color or appearance.

Figure 8 showcases two innovative prompting techniques for generating object descriptions. On the left, we demonstrate how combining an original image with its masked version can effectively prompt GPT-4 to generate detailed descriptions. The right side of Figure 10 highlights the application of the SOTA 'Set of Mark' (SoM[1] [60])

Table 7. Rollback stats on the validation split of three RIS datasets.

Dataset	RefCOCO	RefCOCO+	G-Ref
Num expressions	10,834	10,758	4,896
Rollback rate(%)	10.7	10.8	3.1
mIoU w/o rollback	76.23	68.45	69.01
mIoU w. rollback	78.16	70.12	69.73

[1] https://github.com/microsoft/SoM

Table 8. Universality of VATEX. We conduct experiments to plug-and-play CLIP Prior and MCC in ReLA. † means we run experiment on their official code to get the mIoU score.

Method	RefCOCO	G-Ref
ReLA†	73.16	63.64
ReLA + CLIP Prior	74.32 +1.16	65.76 +2.12
ReLA + MCC	75.46 +1.16	65.12 +1.48
ReLA + CLIP Prior + MCC	76.33 +3.17	67.69 +4.05

Table 9. Fair Backbone Comparison between CRIS, JMCELN, LAVT and VATEX.

Method	Backbone		RefCOCO		
	Visual	Textual	val	testA	testB
CRIS [53]	ResNet-101	CLIP	70.47	73.18	66.10
JMCELN [21]	ResNet-101	CLIP	74.40	77.69	70.43
VATEX (Ours)	ResNet-101	CLIP	**75.66**	**77.88**	**72.36**
LAVT [62]	Swin-B	BERT	74.46	76.89	70.94
LAVT [62]	Swin-B	CLIP	73.15	75.24	70.02
VATEX (Ours)	Swin-B	CLIP	**78.16**	**79.64**	**75.64**

technique to enhance the capability of GPT-4(V) in acquiring deeper knowledge. SoM involves creating masks for each object in the image using SAM, each distinguished by a unique identifier. This marked image then serves as an input for GPT-4V, enabling it to respond to queries necessitating visual grounding with greater accuracy and relevance.

9. Additional Results and Analysis

9.1. Universality of VATEX

VATEX employs CLIP Prior for Object Understanding and Meaning Consistency Constraint for Context Understanding. These two modules can be easily integrated into any DETR-based model (*e.g.* ReLA [34]) for RIS. We took ReLA as a representative work and reproduced the performance of ReLA on the validation sets of the RefCOCO and G-Ref datasets using mIoU metrics. As illustrated in Table 8, VATEX seamlessly integrates into current models, achieving significant performance gains of 3.17% on RefCOCO and 4.05% on G-Ref. This demonstrates the effectiveness of our approach in utilizing Vision-Aware text features for both object understanding and context understanding.

9.2. Additional Comparison on RefCOCO(+/g)

9.2.1 Fair backbone comparison

We have benchmarked our model, VATEX, using the ResNet-101 backbone, aligning it with CRIS and JMCELN for a more equitable comparison, as illustrated in Table 9. This adaptation demonstrates VATEX's superior performance, achieving a 1.26% improvement on RefCOCO val and a significant 1.93% on RefCOCO testB over the current state-of-the-art methods.

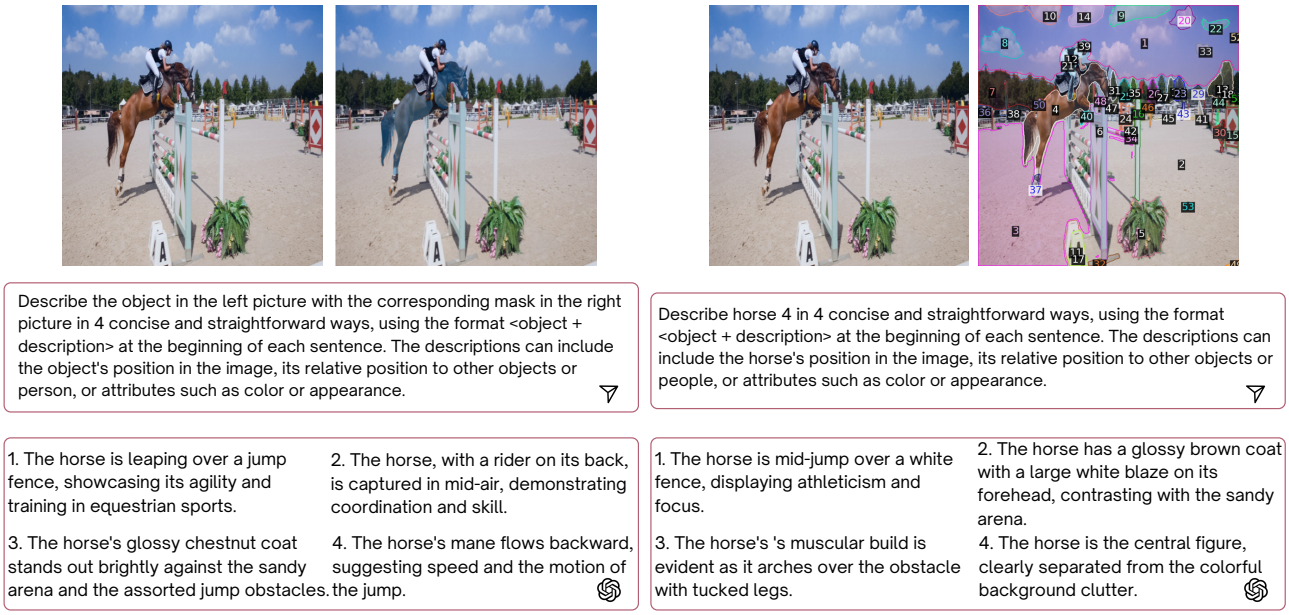
Further, to address comparisons with LAVT, we have experimented with CLIP as the text encoder, adhering to the official repository guidelines. This experiment revealed a performance decline of approximately 1% when substituting BERT with CLIP as the text encoder. This finding underscores the critical importance of using the CLIP Image Encoder together with the CLIP Text Encoder to maintain model performance. A similar trend was observed with ReferFormer, reinforcing our conclusion. Consequently, when compared to LAVT under the fair conditions in backbone, VATEX shows a substantial improvement, outperforming by 5.01%, 4.40%, and 5.62% on RefCOCO val, testA, and testB, respectively. This data confirms the effectiveness of our approach and the importance of consistent backbone usage for fair and accurate performance assessment.

9.2.2 External/Multiple Training dataset

We compare VATEX with other methods in RIS, which used external training data in Table 10. SeqTR [66], RefTR [30], and PolyFormer [36] enhance their performance on the RefCOCO dataset by incorporating external datasets—Visual Genome with 5.4M descriptions across over 33K categories, Flickr30k-entities with 158K descriptions, and the joint dataset RefCOCO(+/g) with 368K descriptions. Their papers indicate that using such external datasets for pretraining can improve performance by 8-10%.

Compared to PolyFormer [36], without using external pretraining dataset, $\text{VATEX}_{\text{RefCOCO}}$ demonstrates superior performance over PolyFormer-B, while $\text{VATEX}_{\text{RefCOCO+}}$ and $\text{VATEX}_{\text{G-Ref}}$ achieve comparable results with [36] while using **42x** and **69x** smaller datasets respectively, with the exception of the RefCOCO+ test B. The performance's gap on RefCOCO+ Test B, which focuses on non-human objects described purely by their appearance (*e.g.* "the porcelain throne," "part of the bed occupied by a black pamphlet"), could be attributed to the varied object categories covered during the pre-training phase with extensive external datasets.

On the otherhand, $\text{VATEX}_{\text{joint}}$ adopts a different strategy. By solely utilizing the RefCOCO(+/g) dataset, which is **16x smaller** than the datasets used by PolyFormer, $\text{VATEX}_{\text{joint}}$ with Swin-B backbone still achieves remarkable results. Specifically, $\text{VATEX}_{\text{joint}}$ outperforms PolyFormer by 4-6% across all benchmarks, setting a new state-of-the-art result on the RefCOCO dataset. UNINEXT [59] and HIPIE [51], while achieving strong results, rely on extensive pretraining and data leakage in finetuning (joint training with COCO for



(a) Prompting with Mask

(b) Prompting with SoM

Figure 8. Example of using GPT-4(V) with different prompting techniques to generate object description.

segmentation while RefCOCO images and annotations are a subset of COCO train split). In contrast, VATEX achieves competitive performance without relying on such extensive pretraining and removes all potential data leaking in the training phase.

9.2.3 Comparison with SOTA foundation models

Table 11 illustrates the quantitative performance between VATEX with generalist foundation models: Grounded-SAM [38] [27], SEEM [69] and X-Decoder [68] in Table 11. For Grounded-SAM, we first use Grounding DINO to extract the bounding box prediction from the text prompt, then we feed that bounding box to SAM to obtain the final segmentation mask. For X-Decoder and SEEM, we directly use the report number on their official github[2] with Focal-L backbones. While VATEX is trained on much smaller dataset sizes and smaller backbones, $\text{VATEX}_{\text{joint}}$ still significantly outperforms Grounded-SAM with 14.34%, 15.65%, and 16.4% improvements on RefCOCO, RefCOCO+ and G-Ref, respectively. Compared with X-Decoder and SAM, which are trained and finetuned on RefCOCO(+/g) datasets, we also outperform them with approximately 2% with VATEX and 7.7% with $\text{VATEX}_{\text{joint}}$.

[2]https://github.com/UX-Decoder/Segment-Everything-Everywhere-All-At-Once/

9.3. Experimental results on Ref-YoutubeVOS and Ref-DAVIS17

The result for Ref-Youtube-VOS dataset is shown in Table 12. As can be seen, our method demonstrates superior performance, setting a new state-of-the-art for referring video object segmentation on the Ref-Youtube-VOS dataset with different backbones. In particular, our approach with the spatial-temporal backbone (e.g., Video-Swin [39]) and pre-trained weights from image dataset achieves the highest $\mathcal{J}\&\mathcal{F}$ score of 65.4% among all other methods on the Ref-Youtube-VOS dataset, including VLT and ReferFormer.

The results for Ref-DAVIS17 are shown in Table 13. Similarly, our approach achieves competitive performance compared to other state-of-the-art methods in referring video object segmentation. Specifically, with backbones ResNet-50, our proposed model outperforms ReferForme and achieves slightly better results than RRVOS. Moreover, our method achieves the best performance among all methods with the Video-Swin-B backbone with a $\mathcal{J}\&\mathcal{F}$ score of 65.4%, which is 3.8% higher than the closest competitor VLT.

9.4. Heatmap of CLIP Prior

To obtain the heatmap result, from the vector of shape $\left(\frac{H}{16} \times \frac{W}{16} + 1, 1\right)$, we remove "CLS" token and reshape it into 2D heatmap of $\frac{H}{16} \times \frac{W}{16}$. For visualization purposes, we resize the original image to 960×960, then pass it through CLIP-Image Encoder, resulting in a high-quality heatmap

Table 10. Quantitative results of referring image segmentation on Ref-COCO, Ref-COCO+, G-Ref datasets with other SOTA methods using external training data. VATEX is trained with Swin-B backbone

Method	External Datasets	RefCOCO			RefCOCO+			G-Ref	
		val	testA	testB	val	testA	testB	val	test
SeqTR [66]	Visual Genome (5.4M) &	71.7	73.31	69.82	63.04	66.73	58.97	64.69	65.74
RefTR [30]	Flickr30k-entities (158K) &	74.34	76.77	70.87	66.75	70.58	59.4	66.63	67.39
PolyFormer-B [36]	RefCOCO(+/g) (368K)	75.96	77.09	73.22	70.65	74.51	64.64	69.36	69.88
UNINEXT-H [59]	Object365 (30M) &	82.2	–	–	72.5	–	–	74.7	–
HIPIE [51]	COCO + RefCOCO(+/g)	**82.6**	–	–	73.0	–	–	75.3	–
VATEX$_{\text{RefCOCO}}$	RefCOCO (142K)	78.16	79.64	75.64	-	-	-	-	-
VATEX$_{\text{RefCOCO+}}$	RefCOCO+ (141K)	-	-	-	70.02	74.41	62.52	-	-
VATEX$_{\text{G-Ref}}$	G-Ref (85K)	-	-	-	-	-	-	69.73	70.58
VATEX$_{\text{joint}}$	RefCOCO(+/g) (368K)	81.53	**82.75**	**79.66**	**74.61**	**78.75**	**68.52**	**75.54**	**76.4**

Figure 9. Our heatmap from CLIP Prior. Naive Implementation means feeding the whole sentence through CLIP Model, without the Main Object Extractor. By reducing the complexity of the text expression, it can be seen that the activation on the object of interest becomes more accurate. Best view in zoom.

Table 11. Quantitative results of referring image segmentation on Ref-COCO, Ref-COCO+, G-Ref validation datasets with SOTA vision foundation models.

Method	RefCOCO	RefCOCO+	G-Ref
Grounded-SAM [38] [27]	67.19	58.96	59.14
X-Decoder [68]	-	-	67.5
SEEM [69]	-	-	67.8
VATEX	78.16	70.02	69.73
VATEX$_{\text{joint}}$	**81.53**	**74.61**	**75.54**

of size 60×60. Notably, we only use a default input size of 224×224 during training. Regarding the quality of the heatmap, Figure 9 demonstrates the comparison between the naive implementation and our prompt-based template. In the 3rd and 7th rows, it is evident that simplifying the sentence and employing prompt templates can aid in distinguishing the target object from the image, resulting in decreased localization errors.

While CLIP Prior excels at localizing objects of interest, it can struggle in complex cases where the expression describes multiple instances within the same category and their relative positions (*e.g.* bottom right of Figure 9). In these situations, the heatmap may encompass all objects within the category rather than the specific referred instances. However, CLIP Prior's core purpose is to narrow down the relevant region, not pinpoint the exact object. Identifying the precise instance will be handled later in the full-text prompt by the Transformer architecture, which can leverage additional context and relationships.

Moreover, CLIP Prior can also help the model in cases when the referring expression contains out-of-vocabulary objects. By transferring the knowledge from CLIP and em-

Table 12. Quantitative comparison with the SOTA on Ref-Youtube-VOS.

Methods	Backbone	Ref-Youtube-VOS		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Train with Image segmentation weight from RefCOCO(+/g)				
ReferFormer [54]	ResNet-50	55.6	54.8	58.4
RR-VOS [32]	ResNet-50	57.3	56.1	58.4
VATEX (Ours)	**ResNet-50**	**58.5**	**57.1**	**59.9**
ReferFormer [54]	Swin-L	62.4	60.8	64.0
VATEX (Ours)	**Swin-L**	**64.2**	**61.4**	**67.0**
ReferFormer [54]	Video-Swin-B	62.9	61.3	64.6
VLT [11]	Video-Swin-B	63.8	61.9	65.6
VATEX (Ours)	**Video-Swin-B**	**65.4**	**63.3**	**67.5**

Table 13. Quantitative comparison with the SOTAs on Ref-DAVIS17 dataset.

Methods	Backbone	Ref-DAVIS17		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
ReferFormer [54]	ResNet-50	58.5	55.8	61.3
RR-VOS [32]	ResNet-50	59.7	57.2	62.4
VATEX (Ours)	ResNet-50	**61.2**	**58.2**	**64.3**
ReferFormer [54]	Video-Swin-B	61.1	58.1	64.1
VLT [11]	Video-Swin-B	61.6	58.9	64.3
VATEX (Ours)	Video-Swin-B	**65.4**	**62.3**	**68.5**

bedding the heatmap into the query initialization, the model can obtain a good segmentation mask based on the cues from CLIP Prior. Figure 10 shows how CLIP Prior heatmap can help the model to localize the object in the early phase, thus improving the model's performance.

CLIP-based model in RIS. Adopting CLIP is a good practice taken by several previous methods, including CRIS, CM-MaskSD, and RIS-CLIP. However, to effectively use the aligned embedding from CLIP to obtain good results in referring segmentation is an open question. For example, although using powerful CLIP as the backbone, the SOTA CLIP-based method RIS-CLIP [26] has a comparable performance with the SOTA Non-CLIP model VG-LAW [47]. To analyze it, we take CRIS [53] as a baseline for CLIP-based model. CRIS directly used the well-aligned embedding space between text and vision for RIS. However, the performance of this work is not good compared to others, as there are two concerns with relying solely on CLIP for referring image segmentation tasks:

1. Frozen CLIP Model. CLIP model, trained on object-centric images, generates visual features focusing on semantic class meanings rather than instance-based details (see bird example in Figure 9). This limits the effectiveness of CLIP for instance-level tasks.

2. Fine-tuning CLIP Model. Fine-tuning the CLIP model risks overfitting on training samples, thereby diminishing its ability to generalize features to novel classes.

Table 14. Quantitative results of referring image segmentation on Ref-COCO, Ref-COCO+, G-Ref validation datasets on CLIP-based and Non-CLIP model.

Method	RefCOCO	RefCOCO+	G-Ref
CLIP-based Model			
CRIS [53]	70.47	62.27	59.87
CM-MaskSD [49]	72.18	64.47	62.67
RIS-CLIP [26]	75.68	69.16	67.62
Ours w/ CLIP Prior	**78.16**	**70.02**	**69.73**
Non-CLIP Model			
LAVT [62]	74.46	65.81	63.34
VG-LAW [47]	75.05	66.61	65.36
Ours w/o CLIP Prior	75.43	67.38	68.12

We found that learning from a visual backbone pre-trained on ImageNet and only utilizing frozen CLIP as a prior gave better performance on both instance-level segmentation and open-vocabulary segmentation nature of RIS task.

In Table 14, for a truly fair comparison, we provide our method w/o CLIP, which achieves 75.43, 67.38, and 68.12 mIoU, and we still outperform the SOTA LAVT (74.46, 65.81, and 63.34) and VG-LAW (75.05, 66.61 and 65.36) on RefCOCO(+/g) in the same setting.

9.5. Full ablation study

Table 15 presents an ablation study conducted on the validation set of RefCOCO and Ref-Youtube-VOS, evaluating the mIoU (mean Intersection over Union) and $\mathcal{J}\&\mathcal{F}$, respectively of different model configurations. The study explores the impact of three components: CLIP Prior, CMD (Contextual Multimodal Decoder), and MCC (Meaning Consistency Constraint).

The first row represents the baseline model with none of the studied components incorporated. The mIoU for this configuration is 70.42% mIoU and 59.8 $\mathcal{J}\&\mathcal{F}$. In rows 2 to 4, the ablation study reveals that incorporating independently the CLIP Prior alone (row 2) and CMD (row 3) both contribute positively to the mIoU on the RefCOCO and $\mathcal{J}\&\mathcal{F}$ on Ref-YoutubeVOS validation set with an improvement of $1.53\%, 2.76\%$ mIoU and $1.7\%, 2.1\%$ $\mathcal{J}\&\mathcal{F}$,

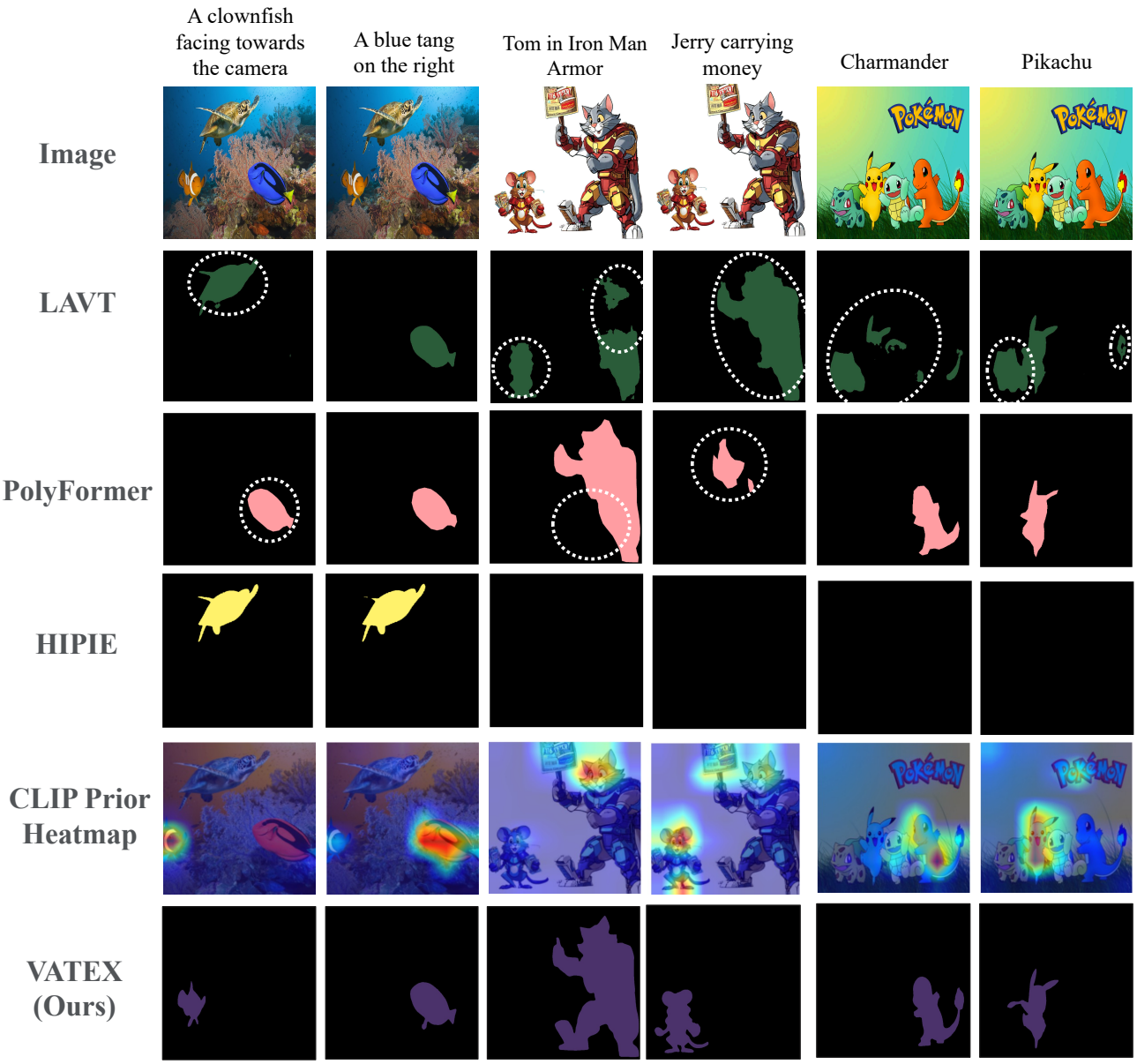


Figure 10. Comparison between VATEX with state-of-the-art methods on challenging out-of-vocabulary cases in referring image segmentation. LAVT's pixel-based approach results in imprecise masks with irrelevant pixel activation. PolyFormer, while creating instance-based masks, struggles with hard cases like "clownfish" or "Jerry" due to limited recognition of unfamiliar objects. HIPIE [51] fails completely due to its constrained pretraining on 365 categories from Objects365. Its high performance on RefCOCO may stem from overfitting and potential data leakage when joint training with COCO. In contrast, VATEX successfully segments correct objects in these difficult vocabulary situations by leveraging the CLIP Prior heatmap. This demonstrates VATEX's superior generalization to unseen objects and complex expressions, highlighting its effectiveness in real-world referring image segmentation tasks.

whereas the introduction of the Meaning Consistency Constraint (MCC) alone (row 4) leads to a modest increase (only 0.30% mIoU and 0.4 $\mathcal{J}\&\mathcal{F}$), emphasizing the individual significance of each component in enhancing model performance. Although MCC alone has a modest impact, when combined with the CMD in row 7, there is a notable improvement of 4.7% (mIoU of 75.1) and 3.3% ($\mathcal{J}\&\mathcal{F}$ of 63.1). This synergy demonstrates that while MCC alone may not perform exceptionally, its collaboration with CMD effectively enhances model performance, aligning with our approach of leveraging enriched text features conditioned by visual information for improved mutual interaction. The final row represents the model with all components (CLIP Prior, CMD, and MCC) combined, achieving the highest

Table 15. Ablation Study on the validation set of RefCOCO (mIoU) and Ref-Youtube-VOS ($\mathcal{J}\&\mathcal{F}$).

	CLIP Prior	CMD	MCC	RefCOCO	Ref-Youtube-VOS
1	-	-	-	70.42	59.8
2	✓	-	-	71.95 +1.53	61.5 +1.7
3	-	✓	-	73.18 +2.76	61.9 +2.1
4	-	-	✓	70.70 +0.30	60.2 +0.4
5	✓	✓	-	75.12 +4.72	63.1 +3.3
6	✓	-	✓	72.14 +1.74	61.3 +1.5
7	-	✓	✓	75.43 +5.01	63.6 +3.8
8	✓	✓	✓	78.16 +7.74	65.4 +5.6

Table 16. Ablation on the number of queries.

Number of queries	1	3	5	10	20	50
RefCOCO	77.23	77.84	78.16	78.02	78.11	77.91

mIoU of 78.16 (+7.74) and $\mathcal{J}\&\mathcal{F}$ of 65.4 (+5.6).

Table 16 presents the impact of varying query numbers on VATEX's performance for the RefCOCO dataset. The results show that while a single query (N=1) achieves a respectable 77.23% mIoU, increasing the number of queries generally improves performance. The optimal performance is achieved with 5 queries, yielding 78.16% mIoU, while the performance slightly decreases for query numbers above 5 (78.02% for 10, 78.11% for 20, and 77.91% for 50 queries). The performance pattern is consistent with ReferFormer [54]'s findings.

9.6. Effect of MCC on Object segmentation mask.

To validate the effectiveness of our proposed MCC module, we propose to use a new Object-centric Intersection over Union (Oc-IoU) score. Unlike mIoU, which averages the overlap and union area for all segmentation predictions within the **same image**, Oc-IoU measures the overlap and union area between the ground truth and all segmentation predictions for the **same object** across different expressions, then averages these values across *all objects in the dataset*. This metric provides an evaluation of the consistency and accuracy of segmentation results across various expressions.

Table 17 provides the comparisons between our method and the state-of-the-art method LAVT in Oc-IoU on the validation set of three RIS benchmarks. As can be seen, our method outperforms LAVT in all three datasets. Comparing the last two rows of Table 17, we can see that the MCC helps the model, especially CMD to enhance mutual information between textual and visual features to further provide more consistent and accurate segmentation. These results underscore the compelling efficacy of our Meaning Consistency Constraint in resolving language ambiguities, thus improving the segmentation performance.

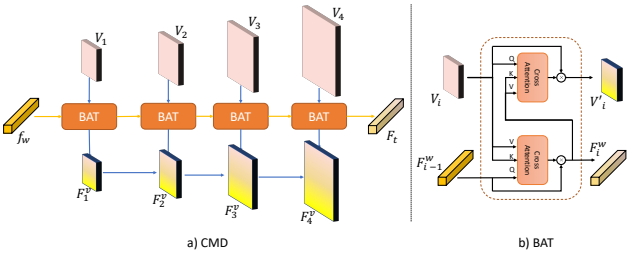


Figure 11. The architecture of Contextual Multimodal Decoder.

Table 17. Performance comparison between LAVT and VATEX on Oc-IoU metric.

Method	RefCOCO	RefCOCO+	G-Ref
LAVT [62]	62.51	50.79	56.01
Ours w/o MCC	66.42	54.92	59.25
Ours	**68.20**	**57.38**	**61.69**

9.7. Archiecture Figure of CMD

For a robust use of visual and text features in subsequent steps, we propose to fuse visual and text features using a Contextual Multimodal Decoder (CMD), which is designed to produce multi-scale text-guided visual feature maps while enhancing contextual information from the image into word-level text features in a hierarchical design as shown in Figure 11. The process on each level of CMD is achieved by a Bi-directional Attention Transfer(BAT), which incorporates two cross-attention modules.

9.8. Runtime and Computational Comparison of VATEX

We report the inference time in FPS and the number of parameters among VATEX, PolyFormer, and LAVT in Table 18. FPS is measured on an NVIDIA RTX 3090 with a batch size of 1 by taking the average runtime on the entire RefCOCO validation set.

Table 18. Comparison in inference time and parameters on the validation set of RefCOCO dataset.

Method	mIoU	FPS	#params	#trainable params
LAVT	74.46	13	217M	217M
PolyFormer	75.96	3.5	295M	295M
VATEX(**Ours**)	78.16	11	251M	165M

Figure 12. Qualitative results of VATEX according to different language expressions for each image on the validation split of G-Ref.

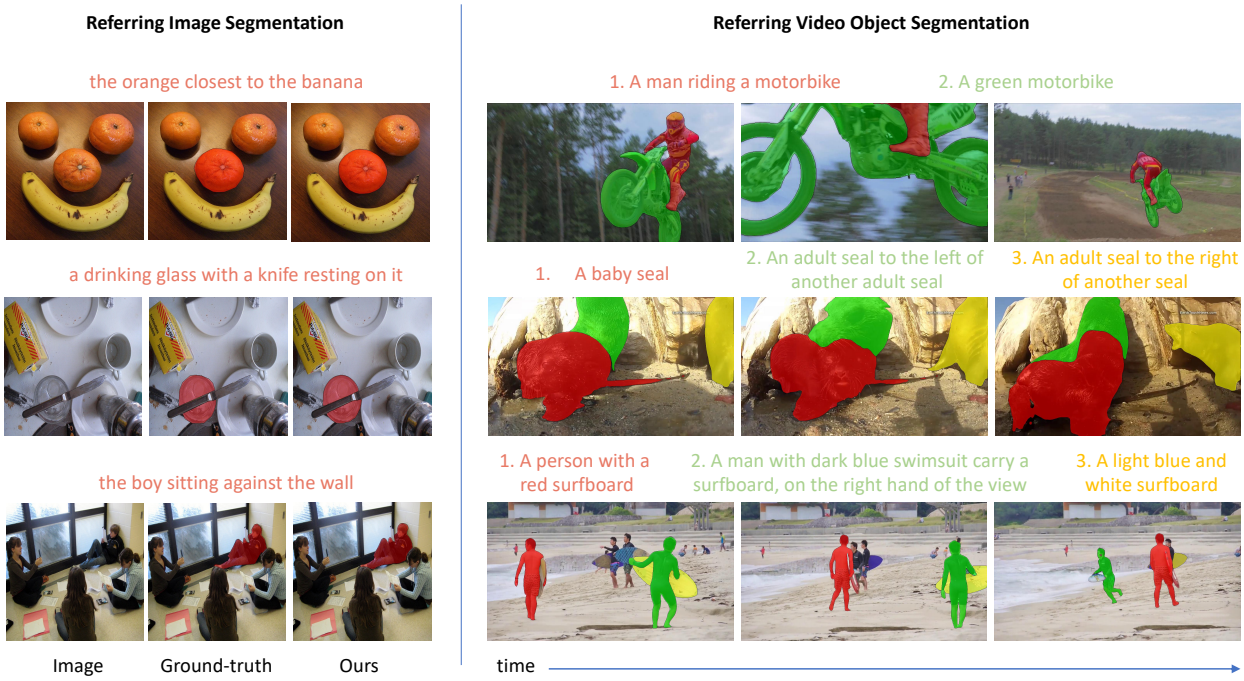


Figure 13. Visualization of VATEX's results. VATEX performs well in complex scenarios such as rapidly changing (*motorbike*), and distinguishing from multiple highly similar objects (*people, seal*). The last row of the video results shows a failure case: PDF segments the wrong man in the last column who has similar attributes when the correct one(green) disappears in the video sequences. Best viewed in color.

9.9. Additional Visual Results

In Figure 12 and Figure 13, we present additional visualization results for our approach. These results demonstrate that VATEX can successfully segment referred objects in a variety of scenarios, including complex expressions or scenes containing multiple similar objects or rapidly changing shapes. To further illustrate our method's capabilities, we have also created a video demo that compares our approach to ReferFormer on Ref-Youtube-VOS. This video demo is provided as an attachment.