Improving Referring Image Segmentation using Vision-Aware Text Features

Anonymous ACL submission

Abstract

Referring image segmentation is a challenging task in computer vision and natural language processing that involves generating pixel-wise segmentation masks based on natural language descriptions. Existing methods have relied mostly on visual features to generate the segmentation masks while treating text features as supporting components. This over-reliance on visual features can lead to suboptimal results, especially in complex scenarios where natural language expressions are ambiguous or context-dependent. To overcome these challenges, we present a novel framework that aims to improve referring image segmentation by utilizing vision-aware text features. Our method involves using the shared embedding of text and image provided by CLIP to derive a CLIP Prior that integrates a visual heatmap with text descriptions, which can be used as the initial query in a transformer architecture for the segmentation task. By observing that our text variations should all refer to the same object in an image, we enforce feature similarity between text variations referring to the same visual input by two components: a novel Contextual Multimodal Decoder that turns text features into vision-aware text embeddings, and a Meaning Consistency Constraint to ensure further the coherent interpretation of language expressions in image analysis. Our method achieves a significant and consistent performance improvement on three benchmark datasets RefCOCO, RefCOCO+ and G-Ref. Code will be released upon acceptance.

1 Introduction

Referring image segmentation (RIS) is an emerging new task in computer vision that aims at predicting pixel-wise segmentation of visual objects in images from natural language cues. Compared to traditional segmentation (Chen et al., 2017; Xie et al., 2021; Wang et al., 2021a,b; Yang et al., 2022a; Heo et al., 2022; Wu et al., 2022b), RIS allows

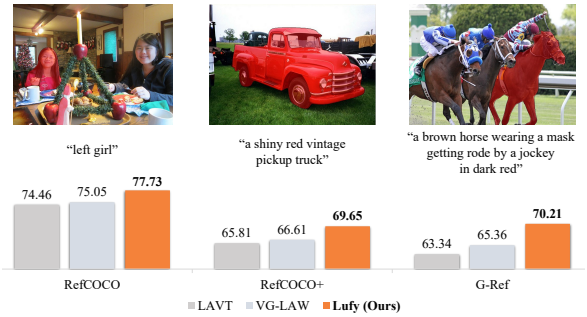

Figure 1: Upper row: visualization of Lufy on three RIS benchmarks. Lower row: Lufy outperforms previous state-of-the-art methods LAVT and VG-LAW, achieving 2-5% gains by utilizing vision-aware text features.

user-friendly interactions via text prompts to select and control the segmentation results, which is useful in various applications, especially interactive applications for image editing and robotics.

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 and context-dependent. Early alignment approaches (Li et al., 2018; Ye et al., 2019; Luo et al., 2020; Feng et al., 2021a) in RIS either used bottom-up methods, merging vision and language features in early fusion and using an FCN as a decoder to produce object masks, or top-down methods, which first identify objects in the image and use the expression as the grounding criterion to select the best-matched result. Recent approaches (Yang et al., 2022b; Ding et al., 2022a; Liu et al., 2023) are based on transformers that learn the interaction between visual-text modalities followed by a standard encoder-decoder process to produce pixel-level segmentation results. These approaches relied mainly on visual features as key and text features are used to enhance visual features. However, we argue that both visual and linguistic features should be considered bidirectionally so that visual features can

also be used to empower text features, resulting in vision-aware text features.

This paper focuses on the investigation of the role of vision-aware text features in improving the performance of RIS. Particularly, we build our method on a transformer architecture for pixel-level segmentation that involves three main components to achieve bidirectional interaction between visual and linguistic features. First, we proposed the idea of integrating visual cues into text features in the query initialization process, facilitated by the visual-text embedding provided by CLIP (Radford et al., 2021). We propose using the alignment between CLIP-Image and CLIP-Text embeddings to generate a visual heatmap that can be subsequently used as an initial mask for our model. By integrating this heatmap into the original text feature, we combine both linguistic and spatial information cues during the query initialization phase. This alignment allows our model to precisely localize the referred object and generate a more accurate segmentation mask.

To further exploit the interaction between visual-text modalities, especially the vision-to-language interaction, we introduce a Contextual Multimodal Decoder, which aims to enhance linguistic features by the contextual information obtained from the visual features as well as to bring the semantic-aware textual information back to visual features in a hierarchical architecture. Finally, given an image, we observe that it is necessary to have a meaningful and consistent understanding of different language expressions when referring to the same instance. To achieve this, we propose the Meaning Consistency Constraint that involves a sentence-level contrastive loss function to enhance the model's ability to grasp the visual information embedded into language expressions. By enforcing this consistency across different expressions referring to the same instance in an image, our approach empowers the model to achieve a more profound and coherent understanding of the visual content of the input.

We present our novel framework for RIS, namely Vision-aware **L**ing**U**istic **F**eatures and Meaning Consistenc**Y** (**Lufy**), which achieves superior performance on all splits of the RefCOCO, RefCOCO+, and G-Ref datasets, surpassing the current state of the art for each datasets by $2-5\%$. We also achieve state-of-the-art performance on both the validation split of Ref-YouTube-VOS and Ref-DAVIS 2017 with **65.4** $\mathcal{J}\&\mathcal{F}$. Figure 1 demonstrates the differences in the complexity of the query sentences between three RIS benchmark datasets and the quantitative result of our proposed method compared to SOTA methods for RIS, which is LAVT (Yang et al., 2022b) and VG-LAW (Su et al., 2023). In summary, our key contributions are as follows.

- We propose a novel transformer framework for referring image segmentation, specially designed to utilize vision-aware text features.
- We present a CLIP-based prior, a Contextual Multimodal Decoder, and a Meaning Consistency Constraint to address the complexities of cross-modal understanding and language ambiguities.
- Through extensive experiments on three referring image segmentation benchmarks (RefCOCO, RefCOCO+, G-Ref) and two referring video segmentation benchmarks (Ref-YouTube-VOS, Ref-DAVIS17), our approach sets the new state of the art.

2 Related Work

Referring image segmentation (Hu et al., 2016) aims to generate pixel-wise segmentation masks for referred objects in images given a natural language expression. Early works (Liu et al., 2019; Huang et al., 2020; Feng et al., 2021b) 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 (Li and Sigal, 2021; seq, 2022; Yang et al., 2022b; Ding et al., 2021, 2022b; Botach et al., 2022), 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 (Su et al., 2023) utilizes language-adaptive weights to dynamically adjust the visual backbone, enabling expression-specific feature extraction for better mask prediction while JMCELN (Huang and Satoh, 2023) learns contextual embeddings and progressively aligns predictions for more accurate image segmentation. PolyFormer (Liu et al., 2023), on the other hand, treats this task as a sequential polygon generation process.

Query Initialization The DETR (DEtection TRansformer) framework (Carion et al., 2020) has

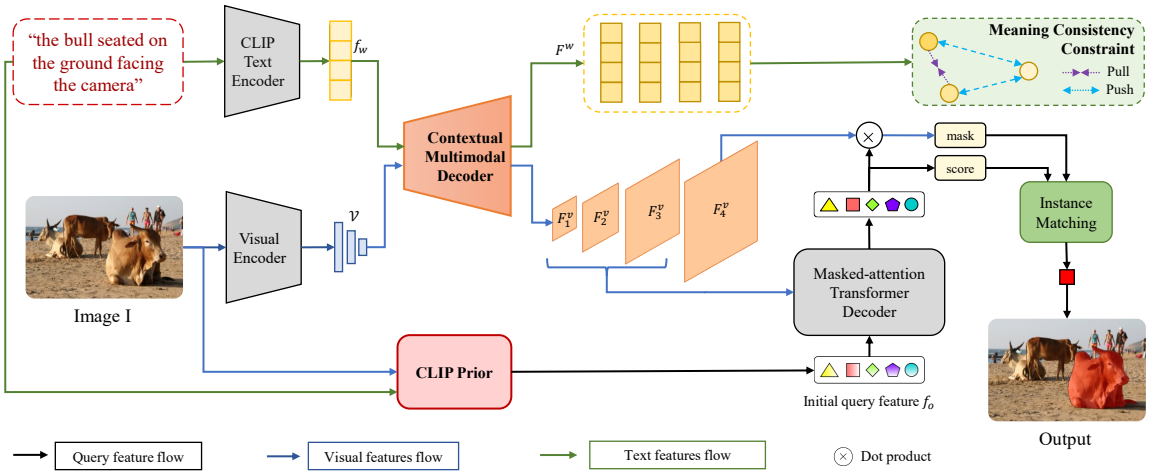

Figure 2: The overview of Lufy. The model takes an image and a language expression as input. The CLIP Prior module is introduced to localize the object of interest based on the text expression and initialize object queries for decoding. Meanwhile, Contextual Multimodal Decoder(CMD) enhances mutual information of both visual and linguistic features in an FPN-like architecture. We further introduce a Meaning Consistency Constraint to control the sentence-level vision-aware text feature space of different expressions. Finally, the enhanced object queries and visual features from the CMD contribute to segmentation mask output.

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 (Zhu et al., 2020) proposes a deformable transformer architecture to refine object queries, while DAB-DETR (Liu et al., 2022a) directly uses the bounding box coordinate in the image to improve query initialization. In the field of referring segmentation, ReferFormer (Wu et al., 2022a) 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 (Chen et al., 2020a,b; He et al., 2020; Wu et al., 2018), enhancing model performance by distinguishing similarities and differences in visual and textual data. CLIP (Radford et al., 2021) employed a contrastive loss on an extensive image-text dataset. CRIS (Wang et al., 2022) leveraged text and pixel-level contrastive learning while VLT (Ding et al., 2022a) applied masked contrastive learning to refine visual features across diverse expressions. Unlike previous approaches (Wang et al., 2022; Ding et al., 2022a) 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

Our framework is constructed by three main components, as demonstrated in Figure 2. First, the image and language expression are initially processed through the Visual Encoder and Text Encoder. We propose a CLIP Prior module to generate a visual heatmap that localizes the object of interest from the text expression, which can be subsequently used to initialize the object queries for the transformer-based method (Sec. 3.1). Next, we propose an efficient way of utilizing cross-attention modules to interact between visual-text modalities in a hierarchical architecture via our Contextual Multimodal Decoder (Sec. 3.2). We further adopt a masked-attention transformer decoder (Cheng et al., 2022) to enhance the object queries through multiscale text-guided visual features. Additionally, we observe that given different expressions referring to the same object within an image, it is necessary to drive embeddings of these expressions as close together as feasible, particularly after the mutual enhancement with visual information via the CMD module. To realize this idea, we lever-

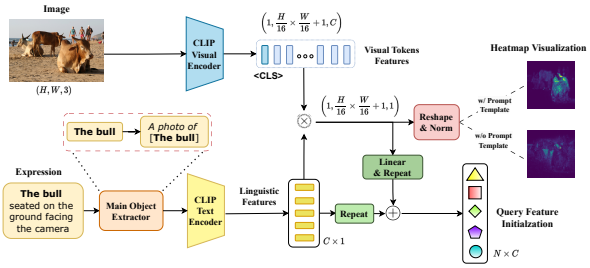

Figure 3: Our CLIP Prior exploits the alignment of CLIP-Image and CLIP-Text embeddings for a better query initialization. Best viewed with zoom.

age sentence-level contrastive learning to enforce a Meaning Consistency Constraint (Sec. 3.3). Finally, the enhanced object queries and the visual features from Contextual Multimodal Decoder are utilized to output segmentation masks (Sec. 3.4).

Mathematically, given an input image $I \in \mathbb{R}^{H \times W \times 3}$ with the size of $H \times W$, we can obtain the multiscale visual feature maps $\mathcal{F}_v = \{f_i^v\}_{i=1}^4, f_i^v \in \mathbb{R}^{H_i \times W_i \times C_i}$ 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 f_i^v. 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

To accurately locate the object of interest in the image from text expression, it is essential to identify the visual region that is the most closely related to the expression. In previous works, query-based methods typically convert textual information from the natural language into object queries and learn target object representation through multi-modal transformers, e.g., using languages as queries (Wu et al., 2022a) and query generation through cross-modal association (Ding et al., 2021). Our goal is to leverage the exceptional alignment between visual and text features in CLIP (Radford et al., 2021) model to generate the heatmap and embed that positional prior in the query.

However, a naive implementation of this idea does not work well. Firstly, when our text expression is input into the CLIP Text Encoder, we obtain a heatmap with inaccurate localization due to the complexity of the input sentence. Additionally, the heatmap obtained from CLIP often contains spot activations in irregular shapes across different regions, which degrades the heatmap's quality. One possible reason for this is that CLIP is trained to map textual descriptions to visual features, enabling it to identify various objects, scenes, and visual concepts. Due to the diversity of these visual concepts, only a few specific features are activated for a particular class or category, while the majority of features remain inactive for numerous other classes. The abundance of non-activated features results in interference and noise in the similarity map between an image and text query. An example is provided in Figure 3 with the input prompt "The bull seated on the ground facing the camera". The heatmap, when not using a prompt template, generates *random masks* covering irrelevant regions in the image with spot activation, which could potentially cause errors for our model. Please refer to the Appendix B.2 for more visualization results.

To address this issue, we first extract the main object (or noun phrase) from the expression (e.g. the bull) in order to obtain a simplified text expression. We assume that a referring expression for segmentation is an object-centric sentence so that it is natural to extract the main object. We then convert the complex referring expression to a simple template-based sentence before passing it to the CLIP Text Encoder. In our implementation, we use "A Photo of [Object]" as our template as it is a common way for a language model to describe an object (Zhou et al., 2022), 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 $\mathcal{F}_{\text{CLIP-Image}} \in \mathbb{R}^{(\frac{H}{16} \times \frac{W}{16} + 1) \times C}$.

While this template approach can efficiently localize regions of interest, particularly for multiple instances of the same class within an image, 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 Transformer framework.

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

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

We embed the positional prior from CLIP 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 object query feature f_o with N queries, each with C-dimension, for embedding the positional prior and text information into the query feature.

3.2 Contextual Multimodal Decoder

For a robust use of visual and text features in subsequent steps, we propose Contextual Multimodal Decoder (CMD) 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 Appendix B.4. Specifically, our CMD is based on a feature pyramid network architecture (Lin et al., 2017), 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.

To enhance the semantic information as well as capture the intra-frame relationships of visual features before being processed by the Contextual Multimodal Decoder, the multi-scale feature maps $\mathcal{F}_v$ are first passed through a stack of Deformable Transformer Encoder layers (Zhu et al., 2021) to produce fine-grained features of the input, represented by $\mathcal{V} = \{V_i\}_{i=1}^4$, where $V_i \in \mathbb{R}^{H_i \times W_i \times C}$ followed (Cheng et al., 2022).

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

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

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

After obtaining F_i^w, we use it as the key and value and V_i as the query in another multi-head cross-attention module to generate features V_i' to

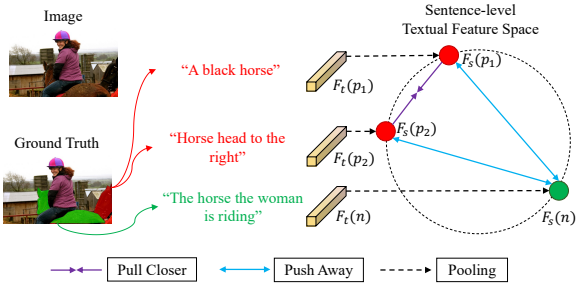

Figure 4: Illustration of Meaning Consistency Constraint. Vision-aware text embeddings of different expressions are made contrastive in sentence-level feature space. Embeddings referring to the same target object are narrowed down while pushing others far away.

reinforce the alignment between the visual and text modalities. We fuse V_i' 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:

$$V_i' = \text{MHA}(V_i, F_i^w, F_i^w) \cdot V_i, \quad (3)$$

$$F_i^v = \text{Conv2d}(V_i' + \text{Ups}(F_{i-1}^v)), \quad (4)$$

where Conv2d() is the 2D convolutional layer, and 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. The final vision-aware text feature is therefore $F_t = F_4^w$.

3.3 Meaning Consistency Constraint

In an image, it is common that each object can be referred to by different text expressions. However, in previous works, the relationship between these expressions has been under-exploited.

To tackle this problem, we propose a sentence-level contrastive loss named Meaning Consistency Constraint (MCC) to learn meaningful and discriminative representations for different expressions while consistently pulling sentences referring to the same object close to each other. Unlike previous approaches (Ding et al., 2022a; Wang et al., 2022), 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 also lead to the improvement of visual features due to the bidirectional attention and hierarchical design of CMD module.

Our contrastive learning is designed as follows. During training, we sample triplets of text expressions that contain two sentences referring to the

Method	Backbone		RefCOCO			RefCOCO+			G-Ref	
	Visual	Textual	val	testA	testB	val	testA	testB	val	test
VLT (Ding et al., 2021)	Darknet-53	Bi-GRU	65.65	68.29	62.73	55.50	59.20	49.36	52.090	56.65
CRIS (Wang et al., 2022)	ResNet-101	CLIP	70.47	73.18	66.10	62.27	68.08	53.68	59.87	60.36
CM-MaskSD (Wang et al., 2023)	CLIP-ViT-B	CLIP	72.18	75.21	67.91	64.47	69.29	56.55	62.67	62.69
LAVT (Yang et al., 2022b)	Swin-B	BERT	74.46	76.89	70.94	65.81	70.97	59.23	63.34	63.62
JMCELN (Huang and Satoh, 2023)	ResNet-101	CLIP	74.40	77.69	70.43	66.99	72.69	57.34	64.08	64.99
VG-LAW (Su et al., 2023)	ViT-B	BERT	75.05	77.36	71.69	66.61	70.30	58.14	65.36	65.13
SeqTR (seq, 2022)†	Darknet-53	Bi-GRU	71.7	73.31	69.82	63.04	66.73	58.97	64.69	65.74
RefTR (Li and Sigal, 2021)†	ResNet-101	BERT	74.34	76.77	70.87	66.75	70.58	59.4	66.63	67.39
PolyFormer-B (Liu et al., 2023)†	Swin-B	BERT	75.96	77.09	73.22	**70.65**	74.51	**64.64**	69.36	69.88
Lufy (Ours)	Swin-B	CLIP	**77.73**	**79.55**	**74.85**	69.65	**74.66**	61.12	**70.21**	**70.53**

Table 1: Quantitative results of referring image segmentation on Ref-COCO, Ref-COCO+, G-Ref datasets. † means using external training data. The best results are in **bold** and the second best ones are in underlined.

same object as positive samples and one describing another object in the same image as the negative sample. We denote the positive samples by p_1, p_2 and the negative sample by n, respectively.

As shown in Figure 4, we gain the sentence-level feature by averaging the vision-aware word-level textual features:

$$F_s = \text{Avg}(F_t, \text{dim} = 0), \tag{5}$$

where $F_t \in R^{L\times C}, F_s \in R^C$. We adopt the InfoNCE loss (Oord et al., 2018) to ensure that linguistic features referring to the same object converge, while features of different objects diverge:

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

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 (Cheng et al., 2022) to update our initial query feature f_o by using the multi-scale visual features $\{F_i^v\}_{i=1}^3$ and also leveraging the heatmap generated from CLIP Prior as the initial attention mask to obtain the final object queries $F_o \in \mathbb{R}^{N\times C}$. They 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:

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

During inference, the query with the highest probability score is selected as the target object for the final output.

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 via a Hungarian algorithm (Kuhn, 1955) by minimizing the matching cost in terms of probability and segmentation mask (Cheng et al., 2021, 2022).

Training. We supervise our prediction $\hat{y}_\delta$ by three losses. Firstly, the class loss $\mathcal{L}_{cls}$ is binary cross entropy (BCE) loss to supervise the probability of 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:

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

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

4 Experimental Results

4.1 Experiment Setup

We evaluate the performance of our model on three image datasets: RefCOCO (Kazemzadeh et al., 2014), RefCOCO+ (Kazemzadeh et al., 2014), G-Ref (Mao et al., 2016) and further evaluate the performance of our model on two video datasets: Ref-Youtube-VOS (Seo et al., 2020) and Ref-DAVIS17 (Khoreva et al., 2019).

For evaluation metrics, we calculate mIoU to measure the overlap between the predicted segmentation mask and the ground truth mask for image segmentation. We also calculate region similarity $\mathcal{J}$ and contour accuracy $\mathcal{F}$, and their average $\mathcal{J}\&\mathcal{F}$ for video object segmentation.

We use PyTorch to implement our method. During training, we freeze both the CLIP Visual and

Methods	Ref-Youtube-VOS			Ref-DAVIS17		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
ReferFormer (Wu et al., 2022a)	62.9	61.3	67.5	61.1	58.1	64.1
VLT (Ding et al., 2022a)	63.8	61.9	65.6	61.6	58.9	64.3
Lufy (Ours)	**65.4**	**63.3**	**67.5**	**65.4**	**62.3**	**68.5**

Table 2: Quantitative results on Ref-Youtube-VOS and Ref-DAVIS17 using Video-Swin-B backbone.

Text Encoder. 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 Appendix A.

4.2 Main Results

Referring Image Segmentation. We compare our proposed method with the state-of-the-art methods on three common datasets (RefCOCO, RefCOCO+, G-Ref). 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. Particularly, for the RefCOCO dataset, our method outperforms the recent VG-LAW, with mIoU gains of 2.68%, 2.19%, and 3.16% across the three splits, respectively. For the more complex RefCOCO+ dataset, our method outperforms the previous top method, JMCELN, by mIoU margins of 2.66%, 1.97%, and 3.78% in the validation, test A, and test B sets, respectively. The G-Ref dataset presented the most significant challenge, with longer and more complicated language expressions. Despite such, our method achieved remarkable performance improvements of approximate 5% in mIoU on the validation and test sets, compared to the second-best method, VG-LAW in the same setting.

It's noteworthy that methods like SeqTR, RefTR, and PolyFormer leverage pre-training on Referring Expression Comprehension (REC) tasks and joint training with Referring Image Segmentation (RIS) later to achieve significant performance gains. While this multi-task strategy can boost mIoU on RefCOCO(+/g) by 3-10% according to their paper, our method outperforms on RefCOCO and achieves comparable performance to PolyFormer-B on RefCOCO+ and G-Ref, without requiring any additional REC training data.

CLIP Prior	CMD	MCC	RefCOCO (mIoU)
-	-	-	70.4
✓	-	-	71.9 +1.5
-	✓	-	73.1 +2.7
-	✓	✓	75.4 +5.0
✓	✓	✓	**77.7** +7.3

Table 3: Ablation Study on the validation set of RefCOCO.

Referring Video Segmentation. Our model can be extended to video datasets with minor adaptations to handle temporal information. As shown in Table 2, Lufy outperforms the current SOTA methods VLT and ReferFormer on the same Video-Swin-B backbone by 1.6 and 3.8 $\mathcal{J}\&\mathcal{F}$ on the Ref-Youtube-VOS and Ref-DAVIS17 datasets, respectively.

Prompt Template	RefCOCO val (mIoU)
Without Prompt Template	74.34
A Photo of [Object]	77.73
Mean of 80 ImageNet prompts	77.91

Table 4: Analysis of prompt templates in CLIP Prior.

4.3 More Analysis

Ablation Study. We conduct an ablation study on the validation sets of RefCOCO with Swin-B backbone to examine the impact of each proposed component in our model. As shown in Table 3, the combination of CLIP Prior, CMD, and MCC modules results in the best performance, showcasing a remarkable performance increase of up to 7.3% in mIoU on RefCOCO. This outcome unequivocally attests to the remarkable effectiveness of our approach and underscores its significant impact. The full ablation study is shown in the Appendix B.3.

CLIP Prior Analysis. As described in Sec. 3.1, our CLIP Prior relies on a template sentence to convert a complex sentence into a simplified sentence suitable for CLIP. Here we investigate the effects of using different templates and how these affect the final performance. Table 4 demonstrates that the absence of prompt templates leads to a significant deterioration in performance. In this case, CLIP introduces noisy activation on various objects based on their discriminative characteristics within the complex sentence. By using the template "A Photo of [Object]", there is a notable improvement of 3.39% in mIoU. We conducted an additional experiment where we aggregated the text embeddings from 80 ImageNet prompts, which has a very mi-

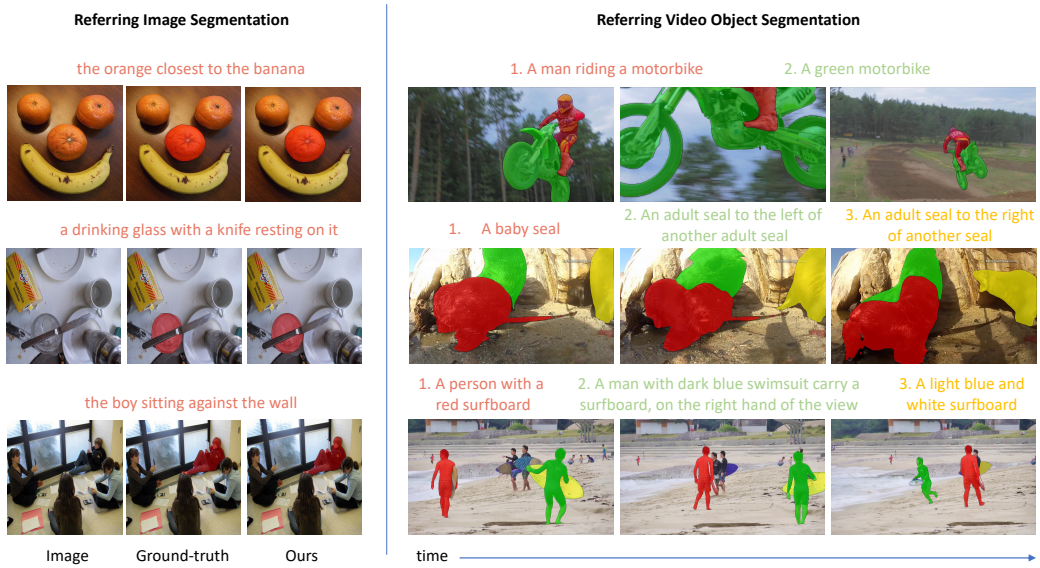

Figure 5: Visualization of Lufy's results. Lufy performs well in complex scenarios such as rapidly changing (*motorbike*), and distinguishing from multiple highly similar objects (*people, seal*). The last row of the video results shows a failure case: Lufy segments the wrong man in the last column who has similar attributes when the correct one(green) disappears in the video sequences. Best viewed in color.

nor performance improvement. This demonstrates that the template "A Photo of [Object]" remains a practical choice.

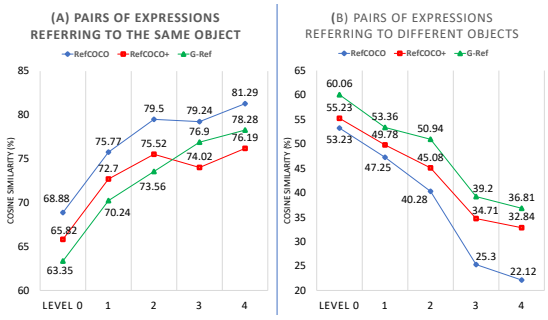

Figure 6: Cosine similarities between sentence-level text features at each CMD level

Effect of CMD on Vision-aware Text Features To study the impact of the Contextual Multimodal Decoder on linguistic features, we quantify the similarity between sentence-level text features at each CMD level. This analysis, depicted in Figure 6, highlights the impact of CMD.

In the context of expressions referring to the same object, the initial similarity of unsupplemented sentence-level features stands at 68.88%, 65.82%, and 63.35% for RefCOCO, RefCOCO+, and G-Ref, respectively. Following just one CMD level, the similarities witness an approximate 7% elevation across all datasets. The culmination of the CMD process results in vision-aware text features yielding similarities of 81.29% in RefCOCO, 78.28% in RefCOCO+, and 76.19% in G-Ref.

Regarding different objects, we compute the similarity between all expressions pertaining to different objects within the same image to showcase the influence of our CMD module on them. As illustrated in Figure 6(b), the text features of these expressions progressively differentiate, transitioning from 53.23%, 55.23%, 60.06% to 22.12%, 32.82%, 36.81% on RefCOCO, RefCOCO+, and G-Ref, respectively.

These results underscore CMD's specialized role in refining text features, crafting an improved representation of vision-aware text features, and then enhancing visual feature representation.

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 or rapidly changing shapes. However, our method fails to segment objects that suddenly disappear as we have not incorporated any special designs for videos.

5 Conclusion

In this work, we proposed Lufy, which addresses the challenge of referring expression segmentation by leveraging natural language descriptions and visual data through a deep multimodal fusion approach. Additionally, Lufy improves the query initialization process with the use of CLIP-Image and CLIP-Text embeddings, achieving state-of-the-art performance on multiple benchmark datasets. Our findings demonstrate the potential of Lufy for applications in image and video search, object recognition, and scene understanding.

6 Limitations

Our method is not without limitations. Particularly, our method has not exploited the interaction and relationship between different objects as well as the alignment between actions and expressions referring to them. Consequently, situations involving counting ("third from left"), indirect descriptions ("a man next to a guy in white shirt"), or actions ("a woman walking to the left") might lead to inaccurate predictions. Dealing with intra-frame object relationships and inter-frame information is vital for future work. It is also of great interest to investigate vision-aware text features with SOTA vision transformer models and SOTA natural language models as well as lifting the referring expression segmentation task to the 3D domain.

References

2022. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision*, pages 598–615. Springer.

Adam Botach, Evgenii Zheltonozhskii, Chaim Baskin, and Chaim Baskin. 2022. 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.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. 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.

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

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

Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. 2020b. Improved baselines with momentum contrastive learning. *ArXiv*, abs/2003.04297.

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

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

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

Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. 2022a. Vlt: Vision-language transformer and query generation for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zihan Ding, Tianrui Hui, Junshi Huang, Xiaoming Wei, Jizhong Han, and Si Liu. 2022b. 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.

Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. 2021a. 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.

Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. 2021b. 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. ISSN: 2575-7075.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. pages 9726–9735.

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

Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. 2016. 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.

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

Ziling Huang and Shin'ichi Satoh. 2023. Referring image segmentation via joint mask contextual embedding learning and progressive alignment network. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7753–7762, Singapore. Association for Computational Linguistics.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. 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. Association for Computational Linguistics.

Anna Khoreva, Anna Rohrbach, and Bernt Schiele. 2019. 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.

Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. 2022. Restr: Convolution-free referring image segmentation using transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18145–18154.

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

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

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

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

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

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

Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. 2023. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18663.

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

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022b. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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

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

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

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

Seonguk Seo, Joon-Young Lee, and Bohyung Han. 2020. 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.

Wei Su, Peihan Miao, Huanzhang Dou, Gaoang Wang, Liang Qiao, Zheyang Li, and Xi Li. 2023. 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.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. 2021a. 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.

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

Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. 2021b. End-to-end video instance segmentation with

transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8741–8750.

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

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

Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. 2022b. 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.

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

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

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

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. 2022b. 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.

Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. 2019. 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.

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

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

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2021. Deformable detr: Deformable transformers for end-to-end object detection.

A Additional Implementation Details

A.1 Datasets

Image datasets. RefCOCO and RefCOCO+ (Kazemzadeh et al., 2014) 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 (Mao et al., 2016), or RefCOCOg, is another well-known dataset with 104,560 language expressions for 54,822 objects in images. The language used in G-Ref is more complex and casual, with longer sentence lengths on average.

Video datasets. Ref-YouTube-VOS (Seo et al., 2020) and Ref-DAVIS17 (Khoreva et al., 2019) 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.

A.2 Training Details

Our model is optimized using AdamW (Loshchilov and Hutter, 2017) optimizer with the initial learning rate of 10^{-5} for the visual encoder and 10^{-4} for the rest. Our model comprises 2 Transformer Encoder and 3 Masked-Attention Transformer Decoder layers, consisting of a total of six Deformable Transformer Decoder layers and nine Masked-Attention Transformer Decoder layers, respectively. We set the number of queries to 5. 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(Khoreva et al., 2019), we directly report the results using the model trained on Ref-YouTube-VOS without finetuning.

B Additional Quantitative Results and Analysis

B.1 Experimental Results on Ref-YoutubeVOS

Methods	Backbone	Ref-Youtube-VOS		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Train with classification weight from Imagenet				
URVOS (Seo et al., 2020)	ResNet-50	47.2	45.3	49.2
LBDT (Ding et al., 2022b)	ResNet-50	49.4	48.2	49.2
MTTR (Botach et al., 2022)	Video-Swin-T	55.3	54.0	56.6
ReferFormer (Wu et al., 2022a)	Video-Swin-T	56.0	54.8	57.3
RR-VOS (Li et al., 2023)	Video-Swin-T	57.1	55.9	58.2
Lufy (Ours)	**Video-Swin-T**	**57.9**	**56.1**	**59.7**
Train with Image segmentation weight from RefCOCO(+/g)				
ReferFormer (Wu et al., 2022a)	ResNet-50	55.6	54.8	58.4
RR-VOS (Li et al., 2023)	ResNet-50	57.3	56.1	58.4
Lufy (Ours)	**ResNet-50**	**58.5**	**57.1**	**59.9**
ReferFormer (Wu et al., 2022a)	Video-Swin-T	59.4	58	60.9
RR-VOS (Li et al., 2023)	Video-Swin-T	61.4	59.6	63.1
Lufy (Ours)	**Video-Swin-T**	**61.7**	**60.0**	**63.4**
ReferFormer (Wu et al., 2022a)	Video-Swin-B	62.9	61.3	64.6
VLT (Ding et al., 2022a)	Video-Swin-B	63.8	61.9	65.6
Lufy (Ours)	**Video-Swin-B**	**65.4**	**63.3**	**67.5**

Table 5: Comparison with the SOTA on Ref-Youtube-VOS.

Methods	Backbone	Ref-DAVIS17		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
ReferFormer (Wu et al., 2022a)	ResNet-50	58.5	55.8	61.3
RR-VOS (Li et al., 2023)	ResNet-50	59.7	57.2	62.4
Lufy (Ours)	ResNet-50	**61.2**	**58.2**	**64.3**
ReferFormer (Wu et al., 2022a)	Video-Swin-B	61.1	58.1	64.1
VLT (Ding et al., 2022a)	Video-Swin-B	61.6	58.9	64.3
Lufy (Ours)	Video-Swin-B	**65.4**	**62.3**	**68.5**

Table 6: Quantitative results of referring video object segmentation on Ref-DAVIS17.

The result for Ref-Youtube-VOS dataset is shown in Table 5. 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. When trained with classification weights obtained from ImageNet, our method with Video-Swin-T achieves a $\mathcal{J}\&\mathcal{F}$ score of 57.9%, which is 0.8% and 1.9% higher than RR-VOS and ReferFormer, respectively. In particular, our approach with the spatial-temporal backbone (e.g., Video-Swin (Liu et al., 2022b)) and pre-trained weights from the image dataset achieves the highest

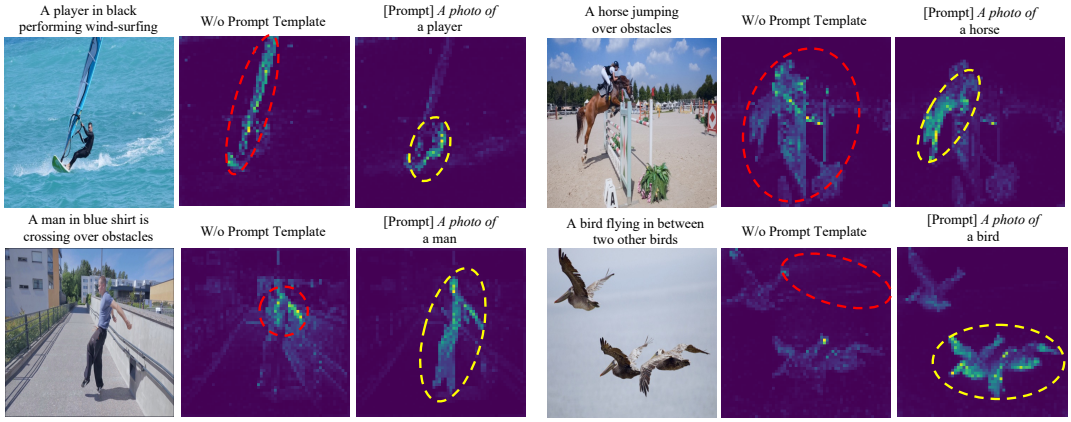

Figure 7: Our heatmap from CLIP Prior. By reducing the complexity of the text expression, it can be seen that the activation on the object of interest becomes more accurate.

$\mathcal{J}\&\mathcal{F}$ 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 6. 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.

B.2 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 purpose, we resize the original image to 960×960, then pass it through CLIP-Image Encoder, resulting in a high-quality heatmap of size 60×60. Notably, we only uses a default input size of 224×224 during training. Regarding the quality of the heatmap, Figure 7 demonstrates the enhanced localization performance achieved when prompt templates are utilized, compared to the naive implementation version of CLIP. Without simplifying the sentence, the heatmap tends to encompass all categories mentioned in the sentence (e.g., obstacles, woman, and blue shirt), or sometimes it may highlight the wrong object (e.g., wind-surfing boat) and also wrong activation at irrelevant regions. In the 3rd and 6th 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 object 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 left of Figure 7). 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.

B.3 Full ablation study

Table 7 presents an ablation study conducted on the validation set of RefCOCO, evaluating the mIoU (mean Intersection over Union) 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 70.4%. 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 validation set with an improvement of 1.5% and 2.7%, whereas the introduction of

N.o	CLIP Prior	CMD	MCC	RefCOCO
1	-	-	-	70.4
2	✓	-	-	71.9 +1.5
3	-	✓	-	73.1 +2.7
4	-	-	✓	70.7 +0.3
5	✓	✓	-	75.1 +4.7
6	✓	-	✓	72.1 +1.7
7	-	✓	✓	75.4 +5.0
8	✓	✓	✓	**77.7** +7.3

Table 7: Ablation Study on the validation set of RefCOCO (mIoU).

the Meaning Consistency Constraint (MCC) alone (row 4) leads to a modest increase (only 0.3%), 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). 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 mIoU of 77.7 (+7.3).

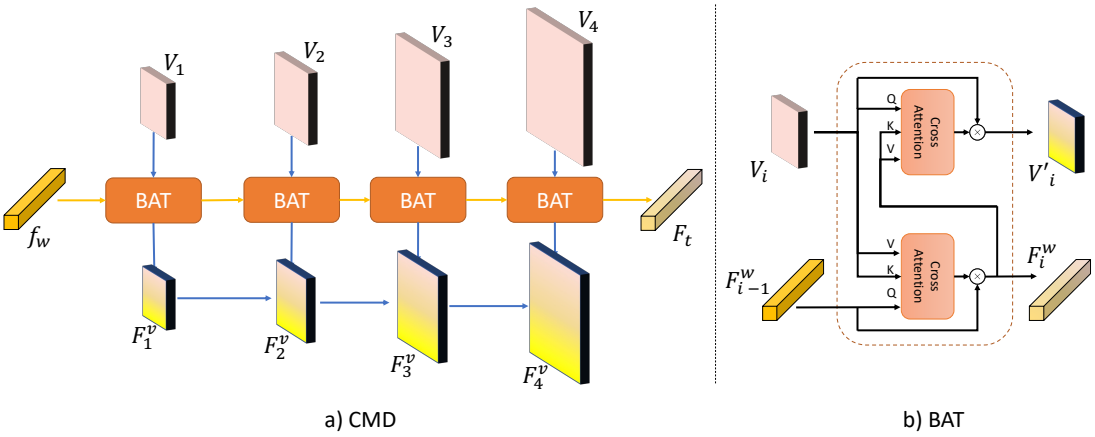

Figure 8: The architecture of Contextual Multimodal Decoder.

B.4 More on Contextual Multimodal Decoder and Meaning Consistency Constraint

B.4.1 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 8. The process on each level of CMD is achieved by a Bi-directional Attention Transfer(BAT), which incorporates two cross-attention modules.

B.5 Meaning Consistency Constraint Analysis

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 use 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. Given a dataset with N objects, and K_i expressions referring

to the i-th object. The Object-centric IoU can be formulated as follows:

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

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

where GT_i and P_i^j denote the ground-truth and the prediction for the j-th expression of the i-th object.

Table 8 provides the comparisons between our method and two previous state-of-the-art methods LAVT and PolyFormer in Oc-IoU. As can be seen, our method outperforms LAVT and PolyFormer on all three datasets. Comparing the last two rows of Table 8, 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.

Method	RefCOCO	RefCOCO+	G-Ref
LAVT (Yang et al., 2022b)	62.5	50.7	56.0
PolyFormer† (Liu et al., 2023)	67.2	56.8	60.9
Ours w/o MCC	66.4	54.9	59.2
Ours	**68.2**	**57.3**	**61.6**

Table 8: Performance on RefCOCO, RefCOCO+, and G-Ref in Oc-IoU metric. † means using external training data.

B.6 Length of language expressions

In Figure 9, we provide a detailed analysis of the performance of our approach based on the length of language expressions. To do this, we followed the approach in (Kim et al., 2022) and split each validation set of the dataset into four groups based on the length of the language expressions (sentence length), with each group being roughly the same size. Our approach Lufy achieved outstanding results, outperforming most previous methods with large margins of over 4% in IoU, except for the group of length 6-7 on RefCOCO, where the difference in performance was only 0.5%. Notably, our approach demonstrated exceptional performance on the G-Ref dataset, surpassing the previous methods (Kim et al., 2022) by a remarkable margin of more than 10% on each sentence length group. These results attest to the effectiveness of our approach in handling the challenges posed by diverse and complex referring expressions in the G-Ref dataset. Additionally, our approach displayed less performance deterioration across the range of sentence lengths on three datasets compared to others. Our findings suggest that our approach accurately identifies objects in complex and diverse visual scenes, rendering it a promising solution for real-world applications where accurate and efficient segmentation is crucial.

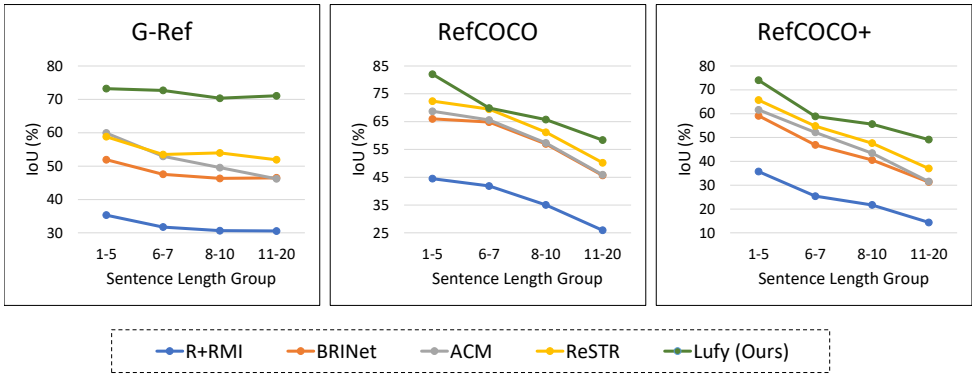

Figure 9: Performance in IoU versus sentence length group on three referring image datasets.

C Additional Visual Results

In Figure 10, we present additional visualization results for our approach. These results demonstrate that Lufy can successfully segment referred objects in a variety of scenarios, including complex expressions and scenes. 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.

Figure 10: Qualitative results of Lufy according to different language expressions for each image on the validation split of G-Ref.