Densely Connected Parameter-Efficient Tuning for Referring Image Segmentation

Jiaqi Huang[1*], Zunnan Xu[1*], Ting Liu[1†], Yong Liu[1], Haonan Han[1], Kehong Yuan[1‡], Xiu Li[1]

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
University Town of Shenzhen, Nanshan District, Shenzhen, Guangdong, P.R. China
{huangjq23,xzn23}@mails.tsinghua.edu.cn, yuankh@sz.tsinghua.edu.cn

Abstract

In the domain of computer vision, Parameter-Efficient Tuning (PET) is increasingly replacing the traditional paradigm of pre-training followed by full fine-tuning. PET is particularly favored for its effectiveness in large foundation models, as it streamlines transfer learning costs and optimizes hardware utilization. However, the current PET methods are mainly designed for single-modal optimization. While some pioneering studies have undertaken preliminary explorations, they still remain at the level of aligned encoders (e.g., CLIP) and lack exploration of misaligned encoders. These methods show sub-optimal performance with misaligned encoders, as they fail to effectively align the multimodal features during fine-tuning. In this paper, we introduce DETRIS, a parameter-efficient tuning framework designed to enhance low-rank visual feature propagation by establishing dense interconnections between each layer and all preceding layers, which enables effective cross-modal feature interaction and adaptation to misaligned encoders. We also suggest using text adapters to improve textual features. Our simple yet efficient approach greatly surpasses state-of-the-art methods with 0.9% to 1.8% backbone parameter updates, evaluated on challenging benchmarks.

Code — https://github.com/jiaqihuang01/DETRIS

1 Introduction

Referring Image Segmentation (RIS) aims to predict the mask of a target object within an image based on a natural language description. Unlike semantic segmentation, which involves assigning a label from a predefined set to each pixel in an image, RIS requires a more nuanced understanding of the language and visual content to identify the described object. The RIS task holds great significance as it effectively bridges the gap between natural language descriptions and fine-grained visual perception (Ji et al. 2024). This capability is crucial for advancing the field of artificial intelligence, particularly in areas such as autonomous systems, image-based retrieval, and human-computer interaction. The complexity of RIS arises from the need to interpret arbitrary context lengths and to comprehend an open-world vocabulary that includes a wide array of object names, attributes, and positional references (Li et al. 2024b). The requirement for precise segmentation of referring objects makes this dense prediction task one of the most challenging tasks in vision language understanding.

In the field of Computer Vision, scaling up foundational models (Radford et al. 2021; Li et al. 2022b; Ma et al. 2022b; Oquab et al. 2023; Fang et al. 2023) is becoming increasingly important. These models leverage large datasets to learn a comprehensive set of visual features. The scaling process not only enhances their ability to discern subtleties in visual data but also significantly boosts their generalization capabilities. With more parameters and exposure to a wider range of data, these models are better equipped to handle diverse and complex visual tasks (Ma et al. 2022a; He et al. 2023, 2024a; Fang et al. 2024; Zhuang et al. 2025), demonstrating robustness that is essential for real-world applications (He et al. 2024b).

However, there often exists a gap between the pre-trained tasks of these models and the specific requirements of downstream applications. Bridging this gap through efficient adaptation presents a formidable challenge. Recent studies (Wang et al. 2022; Ding et al. 2022; Yang et al. 2022; Liu, Ding, and Jiang 2023a; Li et al. 2023) have demonstrated the effectiveness of fine-tuning powerful pre-trained models for referring image segmentation. However, a common challenge is that they typically require full fine-tuning to adapt to dense prediction tasks. This process can lead to the loss of valuable pretraining knowledge, as it involves adjusting a large number of parameters that were previously optimized during the pre-training phase (Kim et al. 2022; Liu, Ding, and Jiang 2023a; Liu et al. 2023a; Wu et al. 2024). Moreover, these approaches maintain a distinct set of fine-tuned parameters of pretrained models for each dataset, which can lead to substantial deployment costs. The problem becomes particularly serious when considering the ever-growing size of pre-trained models, which now include parameters ranging from hundreds of millions to trillions (Li et al. 2022a; Zhou et al. 2022; Chen et al. 2022b; Sun et al. 2023).

Various parameter-efficient tuning methods have been developed to achieve an optimal balance between operational efficiency and model performance (Gao et al. 2021; Chen

*Equal Contribution. Work done as interns at Tencent.

†Work done during Ting Liu's visit at Tsinghua University.

‡Corresponding author

et al. 2022a; Zhou et al. 2022; Wang et al. 2023; Li et al. 2024a; Liu et al. 2024a,c). However, despite these contributions, most existing methods are limited in their scope, predominantly applied to single-modality tasks (Guo, Rush, and Kim 2020; Houlsby et al. 2019; Chen et al. 2022c) or simple classification problems (Gao et al. 2021; Chen et al. 2022a; Zhou et al. 2022). There remains a notable gap in research concerning dense prediction tasks and the nuanced interactions between multiple modalities. Pioneering works such as ETRIS and BarleRIa (Wang et al. 2023) aimed to parameter-efficient fine-tuning CLIP (Radford et al. 2021) for referring image segmentation, but they encountered several limitations: (i) These methods primarily relied on the early-stage fusion of multimodal features from the backbone, missing out on the benefits of more comprehensive global features, which led to suboptimal results. (ii) Furthermore, existing parameter-efficient modules, such as Bridger (Xu et al. 2023) and GST (Wang et al. 2023), are constrained by their limited application of multi-scale modeling. These approach directly fuses information from both modalities, which is insufficient for capturing the full complexity of visual data across different scales.

Our method addresses this question by introducing a simple yet efficient approach that enhances the effectiveness of adapting pre-trained vision-language models through multi-scale modeling and the incorporation of global prior information. In detail, we propose an adapter named Dense Aligner, which can be seamlessly integrated into pre-trained models for dense prediction tasks. There are two customized modules for Dense Aligner: (i) a dense mixture of convolution module designed to capture multi-scale semantic features from the intermediate layers and (ii) a cross-aligner module that facilitates information exchange between visual and textual features. Secondly, we propose incorporating text adapters to enhance the text encoder. We further leverage these enhanced features to improve alignment between visual and linguistic features.

Our framework is built around a dual encoder architecture. Unlike previous PEFT methods that rely on highly aligned encoders (e.g., CLIP), we support DINO (Oquab et al. 2023) as our visual encoder. The reason we chose DINO as our vision backbone is based on several insights: (i) DINO's self-supervised learning approach provides robust generalization and is especially beneficial for dense prediction tasks compared to CLIP (Radford et al. 2021). (ii) The absence of multimodal pre-training in DINO, particularly in visual-text alignment, presents challenges for its direct application on referring image segmentation. This deficiency underscores the essential role of our proposed module in boosting the model's abilities, especially in enhancing visual language alignment and the execution of dense prediction tasks.

Our main contributions are as follows:

- We support the powerful pre-trained model DINO in RIS tasks and provide a parameter-efficient fine-tuning strategy for visual-text alignment that avoids the need for intricate design.
- We propose a simple yet efficient adapter called Dense Aligner that can be seamlessly integrated into the pre-trained backbone to enhance and interact with its intermediate features. This integration improves DINO's alignment with language and enhances its performance on dense prediction tasks.
- Experiments demonstrate that our method greatly surpasses state-of-the-art fully fine-tuned methods in referring image segmentation, with only 0.9% to 1.8% backbone parameter updates.

2 Related Work

Parameter Efficient Tuning (PET) aims to streamline the process of adapting pre-trained models to new tasks with minimal parameter adjustments, making it a practical solution for deploying large models to individual users, particularly in the face of expanding model sizes. Previous PET methods can be mainly divided into three types: (i) updating newly added parameters to the model or input (Houlsby et al. 2019; Li and Liang 2021; Zhou et al. 2022); (ii) sparsely updating a small number of parameters of the model (Guo, Rush, and Kim 2020; Zaken, Ravfogel, and Goldberg 2021; Liu et al. 2024b); (iii) low-rank factorization for the weights to be updated (Hu et al. 2021; Karimi Mahabadi, Henderson, and Ruder 2021; Hao et al. 2023). However, previous works applying PET in computer vision mainly focus on classification and generation tasks. How to efficiently update and transfer the pre-trained knowledge space to dense prediction tasks remains a great challenge. Some pioneering work like ETRIS (Xu et al. 2023) and BarleRIa (Wang et al. 2023) sought to utilize adapters to fine-tune CLIP (Radford et al. 2021) for referring image segmentation. However, their proposed modules like Bridger (Xu et al. 2023) and GST (Wang et al. 2023) are insufficient for capturing the complexity of multi-scale visual features.

Referring Image Segmentation (RIS) aims to segment the target objects referred to by natural language descriptions. It necessitates the models to comprehensively associate diverse visual content and linguistic signals.

The advent of the Transformer model has catalyzed a paradigm transformation of integrating features across diverse modalities by attention mechanism (Yang et al. 2022; Liu, Ding, and Jiang 2023a; Yan et al. 2023; Liu et al. 2023a). Among them, MDETR (Kamath et al. 2021) and VLT (Ding et al. 2022) have demonstrated remarkable performance across various Vision-Language (VL) tasks by integrating multi-modal attention interaction and query representation. Capitalizing on the robust image-text alignment capabilities of CLIP, CRIS (Wang et al. 2022), ETRIS (Xu et al. 2023) and UniLSeg (Liu et al. 2023b) zeroes in on sentence-pixel alignment to harness the wealth of multi-modal correspondences. However, existing methods primarily concentrate on the design of visual-linguistic interactions during the decoding phase, while overlooking the potential to fully excavate the pretrained backbone networks. To this end, we propose DETRIS, a framework that aligns features from different modalities with the assistance of parameter-efficient modules boosting multi-scale comprehensive updating of backbone networks. Compared to existing fully fine-tuning methods, the proposed approach achieves competitive performance while greatly reducing the training

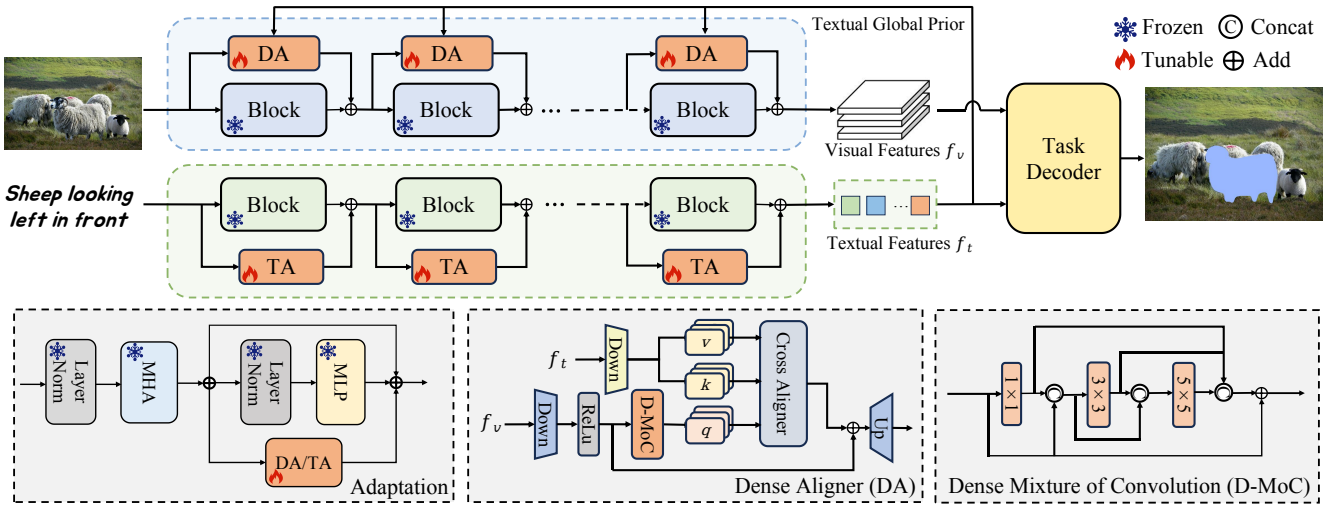

Figure 1: Overall framework of our DETRIS. In the image branch, we utilize Dense Aligner (DA) to facilitate cross-modal and multi-scale modeling of low-rank visual features. This approach incorporates textual global prior information to enhance the visual features f_v. In the text branch, we also use "D-MoC" as our Text Adapters (TA) to obtain the text feature f_t.

computation overhead.

3 Methodology

3.1 Framework Overview

The overall framework of the proposed model is illustrated in Figure 1. Our approach freezes the parameters of the pre-trained backbone, ensuring parameter efficiency. The core of the model's design philosophy is the Dense Aligner, which is intended to facilitate interaction between cross-modal features and inject dense prediction priors into the pre-trained backbone. This is coupled with the incorporation of text adapters in the text encoder to enhance fine-grained image-text alignments.

3.2 Image & Text Feature Extraction

Visual encoder. Our work adapts the distilled DINOv2 with registers (Oquab et al. 2023) as the backbone. This model has been well-pretrained on self-supervised training tasks and is based on ViT-B/14. Specifically, for an input image $I \in R^{H\times W\times 3}$, we utilize DINOv2 enhanced with Dense Aligners to extract image features. We select the outputs of several middle layers and the last layer as hierarchical vision features $f_v^i, i \in [1, 2, 3]$, which are used for subsequent dense prediction modules.

Text encoder. For the input refer expression T, we utilize the pre-trained text transformer of CLIP (Radford et al. 2021) to extract text features. Text features f_t and sentence-level feature f_s are extracted using a CLIP text encoder, augmented with text adapters to guide referring image segmentation.

Considering the substantial number of parameters that the encoder occupies, and to avoid the loss of valuable pre-training knowledge, we freeze all the encoder parameters during our fine-tuning process to efficiently apply it in downstream tasks.

3.3 Local & Global Feature Interaction

As mentioned in Section 1, although DINOv2 has strong generalization capabilities and is more advantageous than CLIP in tasks that rely more on visual abilities, DINOv2 lacks visual-text alignment in the RIS downstream task due to the absence of multimodal pre-training. To address this and enhance the model's multi-scale modeling capability, we designed and utilized Dense Aligners to augment the model's vision backbone. The visual backbone network remains fixed, with training solely focused on the Dense Aligner parameters.

Dense Aligner. As shown in Figure 1, the proposed Dense Aligner differs significantly from previous adapter designs by incorporating a dense mixture of convolution modules. Additionally, a cross-aligner module is integrated between the activation and up-projection layers. This integration enhances the model's ability to extract dense image features and enriches its multimodal fusion capabilities.

In the dense mixture of convolution module, a multi-branch convolutional structure is introduced to effectively model low-rank visual features across multiple scales and to capture multi-scale prior information. The module sequentially applies linear projection and non-linear activation, followed by 1×1, 3×3, and 5×5 convolutional kernels to progressively integrate outputs from previous layers. To maintain efficiency and reduce computational load, 1×1 convolutions are placed before the 3×3 and 5×5 convolutions to compress the channel dimensions.

In this dense mixture module, outputs from smaller kernels serve as inputs for larger kernels, ensuring a detailed stepwise integration of fine and broad features. The 1×1 convolution output feeds into the 3×3 convolution, and the combined outputs of the 1×1 and 3×3 convolutions are then passed to the 5×5 convolution. This method differs from conventional direct merging of multi-scale features by deli-

cately combining fine details with broader contextual information. To preserve the original features, the initial inputs are added back to the final concatenated outputs.

Specifically, for given input image features f_v^l at layer l, this process can be formulated as below:

$$\begin{aligned} F_v^l &= \sigma(\text{Linear}_{\text{down}}(f_v^l)), \\ F_{v1}^l, F_{v2}^l, F_{v3}^l &= \text{D-MoC}(F_v^l) \\ F_{\text{dense}}^l &= (F_{v1}^l, F_{v2}^l, F_{v3}^l) + F_v^l, \end{aligned} \tag{1}$$

where σ denotes non-linear activation function ReLU, $\text{Linear}_{\text{down}}$ denotes a downsampling operation of linear projection. F_{dense}^l is the final dense feature representation obtained by concatenating the features from all branches and adding the initial feature F_v^l. Here, $(,)$ represents concatenating the features along the dimensional direction.

For the D-MoC operation, the process can be formulated as below:

$$\begin{aligned} F_{v1}^l &= \text{conv}_{1\times1}(F_v^l), \\ F_{v2}^l &= \text{conv}_{3\times3}(F_v^l, F_{v1}^l), \\ F_{v3}^l &= \text{conv}_{5\times5}(F_v^l, F_{v1}^l, F_{v2}^l), \end{aligned} \tag{2}$$

where F_{v1}^l, F_{v2}^l, and F_{v3}^l are the output features after applying respective convolutional operations.

Considering that textual information contains valuable references, we utilize it as a global reference prior by integrating it into the vision backbone network via a aligner module. This not only regularizes the visual features but also aligns them better with the global features of the text (denoted as f_t). This process can be formalized as:

$$\begin{aligned} F_{\text{cross}}^l &= \mathcal{F}_{\text{align}}(F_{\text{dense}}^l, f_t) + F_v^l, \\ f_{dc}^l &= \text{Linear}_{\text{up}}(F_{\text{cross}}^l), \end{aligned} \tag{3}$$

where $\mathcal{F}_{\text{align}}$ represents the alignment method, and we found that simple multi-head cross attention is quite effective, F_{cross}^l denotes the fused visual feature, and $\text{Linear}_{\text{up}}$ represents an operation to project the visual features back to get f_{dc}^l. We add the Dense Aligner in parallel to the MLP layer in the transformer block to achieve our adaptation as illustrated in Figure1.

Text Adapter. Acknowledging the disparities between the text features extracted by the CLIP text encoder and the DINO features, we incorporate a text adapter to improve the text encoder for fine-grained alignment of text and visual features. In a manner similar to the Dense Aligner, we employ a Dense mixture of Convolution as our text adapter to better capture multi-scale text information and enhance the alignment between visual and textual features. However, unlike the Dense Aligner, which uses 2D convolutions, our text adapter leverages 1D convolutions specifically designed for processing textual data. This approach ensures that text features are effectively captured and integrated, enabling improved alignment and compatibility with the overall model architecture while optimizing the representation of sequential textual information. Specifically, for given input text features f_t^l at layer l, this process can be formalized as:

$$\begin{aligned} F_t^l &= \text{Linear}_{\text{down}}(f_t^l), \\ F_{\text{relu}}^l &= \text{D-MoC}(F_t^l), \\ f_{\text{w}}^l &= \text{Linear}_{\text{up}}(F_{\text{relu}}^l), \end{aligned} \tag{4}$$

where $\text{Linear}_{\text{down}}$ represents a downsampling linear projection and $\text{Linear}_{\text{up}}$ denotes an upsampling operation to adapt text features back to the original dimension.

3.4 The Referring Image Segmentation Head

To ensure a fair comparison, we follow CRIS (Wang et al. 2022) and ETRIS (Xu et al. 2023), incorporating a learnable referring image head. This head which consists of three main components: a cross-modal neck, a vision-language decoder, and an up-sample projector. These collaborate together to extract the cross-modal feature F_c and the transformed textual feature F_l.

The cross-modal neck takes multiple adapted visual features ($\hat{f}_v^i, i \in [1, 2, 3]$) from three layers of the visual encoder (e.g., the 1/3, 2/3, and the last layer of the backbone) and the adapted textual embeddings $\hat{f}_t$. Specifically, we employ a multi-head cross-attention mechanism ($\mathcal{F}_{\text{MHCA}}$) with convolution to fuse these features, obtaining the fusion features F_f. Subsequently, we concatenate a 2D spatial coordinate feature F_{coord} with F_f and further fuse them using a 3×3 convolution, which can be formalized as:

$$f_c = \text{Conv}([F_f, F_{\text{coord}}]), \tag{5}$$

where $F_f = \mathcal{F}_{\text{MHCA}}(\hat{f}_v^i, \hat{f}_t)$ and f_c denotes the combined cross-modal feature.

The vision-language decoder further merges the composite feature f_c with the textual embeddings $\hat{f}_t$. This fusion process culminates in the generation of multimodal features F_{mm}, encapsulating both visual and linguistic information. Specifically, the decoder consists of three layers, each composed of a multi-head self-attention layer (MHSA), a multi-head cross-attention layer (MHCA), and a feed-forward network. Within each decoder layer, the combined features f_c are fed into the MHSA layer to capture global contextual information. The MHCA layer further facilitates multi-modal interaction by mapping visual features to queries and textual features to keys and values. Following the MHCA layer, an MLP block, along with layer normalization and residual connections, further processes the output features.

An up-sampling projector further transforms the multimodal features F_{mm} and the sentence-level feature f_s to extract the cross-modal feature F_c and the transformed textual feature F_l. f_s is first transformed into F_l through a linear transformation, then split and reshaped into weights and bias, enabling it to function as a Conv2D layer. This Conv2D layer is used to transform the cross-modal representation into the final mask prediction. The overall transformation is achieved using a 4× upsampling followed by convolution and linear projection:

$$\begin{aligned} F_c &= \text{Conv}(\text{UpSample}(F_{mm})), \\ F_l &= \text{Linear}(f_s). \end{aligned} \tag{6}$$

Method	RefCOCO			RefCOCO+			G-Ref			Avg
	val	testA	testB	val	testA	testB	val(u)	test(u)	val(g)	
Traditional Full Fine-tuning										
ReSTR[CVPR 22] (Kim et al. 2022)	67.2	69.3	64.5	55.8	60.4	48.3	54.5	-	54.5	58.8
CRIS[CVPR 22] (Wang et al. 2022)	70.5	73.2	66.1	62.3	68.1	53.7	59.9	60.4	-	63.8
LAVT[CVPR 22] (Yang et al. 2022)	72.7	75.8	68.8	62.1	68.4	55.1	-	-	60.5	64.9
SEEM[NeurIPS 23] (Zou et al. 2023)	-	-	-	-	-	-	65.6	-	-	-
VPD[ICCV 23] (Zhao et al. 2023)	73.5	-	-	63.9	-	-	63.1	-	-	66.8
DMMI[ICCV 23] (Hu et al. 2023)	74.1	77.1	70.2	64.0	69.7	57.0	63.5	64.2	62.0	66.9
ReLA[CVPR 23] (Liu, Ding, and Jiang 2023b)	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0	62.7	67.7
ReLA[CVPR 23] (Liu, Ding, and Jiang 2023a)	73.8	76.5	70.18	66.0	71.0	57.7	65.0	66.0	62.7	67.5
CGFormer[CVPR 23] (Tang et al. 2023)	74.8	77.3	70.6	64.5	71.0	57.1	64.7	65.1	62.5	67.7
LISA-7B[CVPR 24] (Lai et al. 2023)	74.1	76.5	71.1	62.4	67.4	56.5	66.4	68.5	-	67.9
MagNet[CVPR 24] (Chng et al. 2023)	75.2	78.2	71.1	66.2	71.3	58.1	65.4	66.2	63.1	68.3
ReMamber[ECCV 24] (Yang et al. 2024)	74.5	76.7	70.9	65.0	70.8	57.5	63.9	64.0	-	67.9
RISCLIP-B[NAACL 24] (Kim et al. 2024)	75.7	78.0	72.5	69.2	73.5	60.7	67.6	68.0	-	70.6
Parameter Efficient Tuning										
ETRIS[ICCV 23] (Xu et al. 2023)	70.5	73.5	66.6	60.1	66.9	50.2	59.8	59.9	57.9	62.8
BarLeRIa[ICLR 24] (Wang et al. 2023)	72.4	75.9	68.3	65.0	70.8	56.9	63.4	63.8	61.6	66.5
DETRIS-B (Ours)	76.0	78.2	73.5	68.9	74.0	61.5	67.9	68.1	65.9	70.4
DETRIS-L (Ours)	**77.3**	**79.0**	**75.2**	**70.8**	**75.3**	**64.7**	**69.3**	**70.2**	**67.9**	**72.2**
With Mixed Training Data										
PolyFormer-L*[CVPR 23] (Liu et al. 2023a)	76.0	78.3	73.3	69.3	74.6	61.9	69.2	70.2	-	71.6
UNINEXT-L*[CVPR 23] (Yan et al. 2023)	80.3	**82.6**	77.8	70.0	74.9	62.6	73.4	73.7	-	74.4
DETRIS-L* (Ours)	**81.0**	81.9	**79.0**	**75.2**	**78.6**	**70.2**	**74.6**	**75.3**	-	**77.2**

Table 1: State-of-the-art comparison of RIS methods and the PET RIS method on RefCOCO/RefCOCO+/G-Ref datasets without using extra data and Mixed RefCOCO dataset, evaluated using the IoU metric. For Mixed RefCOCO datasets, models marked with * are tuned using the mixed RefCOCO/RefCOCO+/G-Ref datasets. The best results are in bold.

3.5 Training Objective

Following CRIS (Wang et al. 2022), we adopt a text-to-visual contrastive loss, denoted as $\mathcal{L}_{\text{con}}$, for the training objective of our model to optimize the alignment between text-derived features and their corresponding visual pixels.

This contrastive loss is designed to both enhance the connection between text features and corresponding visual pixels, and separate these text features from any unrelated visual elements. The text-to-pixel contrastive loss is mathematically articulated in the following manner:

$$\begin{aligned}\mathcal{L}_{\text{con}}^{i}\left(F_c^i, F_l\right) &= \begin{cases} -\log\left(\sigma\left(F_c^i \cdot F_l\right)\right), & i \in \mathcal{P} \\ -\log\left(1-\sigma\left(F_c^i \cdot F_l\right)\right), & i \in \mathcal{N} \end{cases} \\ \mathcal{L}_{\text{con}}\left(F_c, F_l\right) &= \frac{1}{|\mathcal{P}\cup\mathcal{N}|}\sum_{i\in\mathcal{P}\cup\mathcal{N}}\mathcal{L}_{\text{con}}^{i}\left(F_c^i, F_l\right),\end{aligned} \tag{7}$$

where $\mathcal{P}$ and $\mathcal{N}$ denote the class of 1 and 0 in the ground truth, and σ denotes the sigmoid function. The loss thus penalizes incorrect alignments between features and encourages the model to correctly match textual descriptions to their associated visual representations.

4 Experiments

4.1 Datasets

We employ three challenging referring image segmentation benchmarks in our experiments:

- **RefCOCO (Kazemzadeh et al. 2014)** is widely used as a benchmark for referring image segmentation. It comprises 19,994 images annotated with 142,210 referring expressions for 50,000 objects, which have been sourced from the MSCOCO dataset through a two-player game. The dataset is divided into four subsets, consisting of 120,624 training samples, 10,834 validation samples, 5,657 samples for test A, and 5,095 samples for test B, respectively. The average length of the expressions is 3.6 words, and each image contains a minimum of two objects.
- **RefCOCO+ (Kazemzadeh et al. 2014)** dataset consists of 141,564 referring expressions associated with 49,856 objects in 19,992 images. The dataset is divided into four subsets: 120,624 train, 10,758 validation, 5,726 test A, and 4,889 test B samples. Notably, the RefCOCO+ dataset has been constructed to be more challenging than the RefCOCO dataset by excluding certain types of absolute location words.
- **G-Ref (Yu et al. 2016)** comprises 104,560 referring expressions associated with 54,822 objects in 26,711 im-

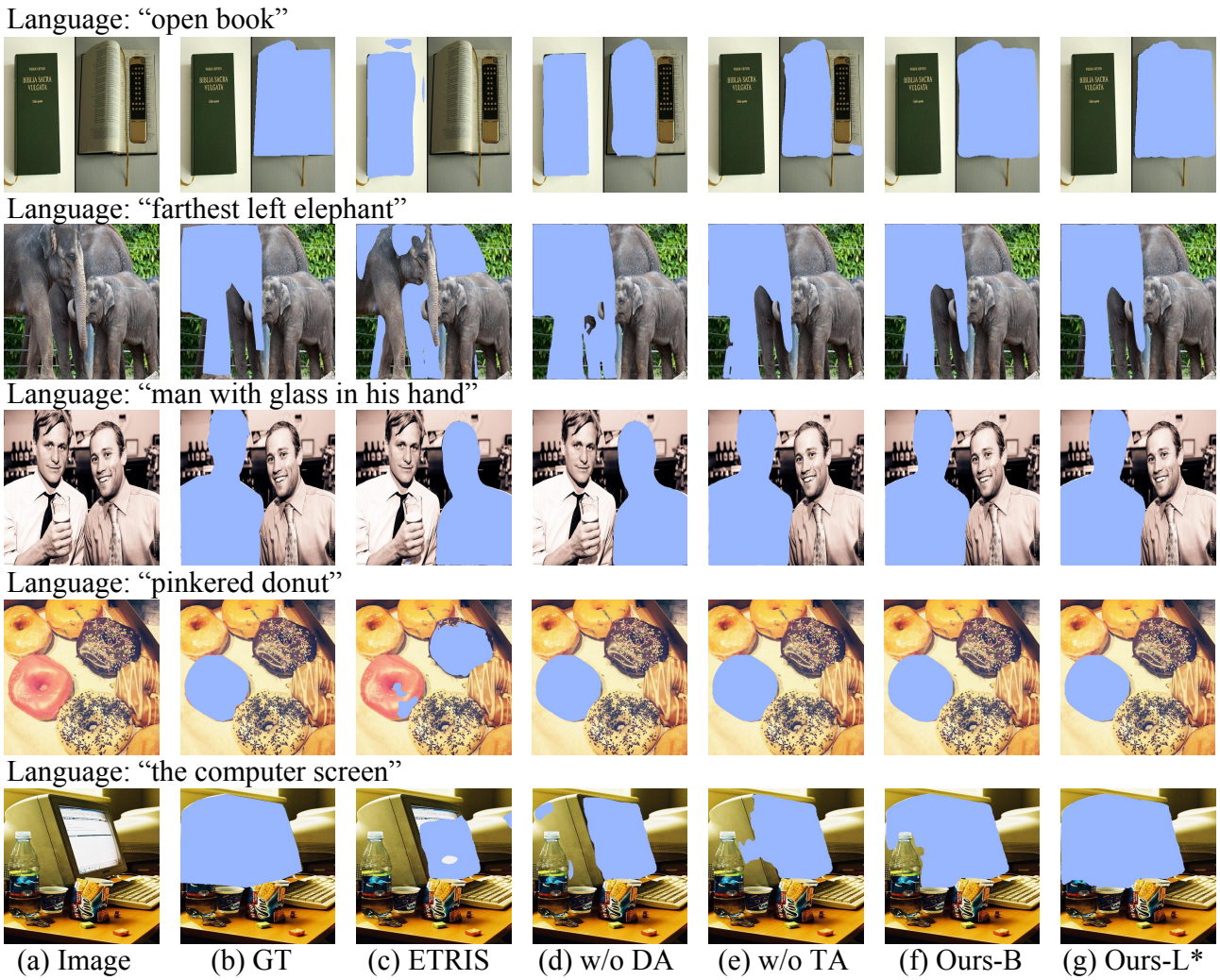

Figure 2: Qualitative results: (a) the input image; (b) the ground truth; (c) ETRIS; (d) DETRIS-B without Dense Aligner; (e) DETRIS-B without Text Adapter; (f) our proposed DETRIS-B; (g) DETRIS-L using mixed datasets.

ages. The expressions in G-Ref were collected from Amazon Mechanical Turk and had an average length of 8.4 words, which included more words related to locations and appearances. We present results for both the Google and UMD partitioning methods for G-Ref.

4.2 Implementation Details

In our experiments, DETRIS-B uses DINOv2-B/14 as the vision backbone, and DETRIS-L uses DINOv2-L/14. Both models employ the CLIP text encoder, with input images resized to 448x448 pixels. The Dense Aligner (dim=128) is applied at layers [1, 3, 5, 7, 9, 11] for DETRIS-B and [2, 6, 10, 14, 18, 22] for DETRIS-L. The Text Adapter (dim=64) is applied at layers [1, 3, 5, 7, 9, 11] in both models. We train the framework end-to-end for 50 epochs using the Adam optimizer. The learning rate starts at 0.0001 and decays by 0.1 at epoch 35. DETRIS-B is trained on 2 A100 GPUs with a batch size of 32, while DETRIS-L uses 4 A100 GPUs with a batch size of 64 and an initial learning rate of 0.0002. Performance is evaluated using mIoU, which measures the intersection-over-union between predicted masks and ground truth, following (Wang et al. 2021).

4.3 Main Results

We compared our DETRIS models with previous RIS methods. As shown in Table 1, DETRIS-B achieves 70.4 IoU and DETRIS-L reaches 72.2, improving by 3.7% and 6.6% over the previous state-of-the-art. DETRIS-L outperforms all methods, especially on the challenging RefCOCO+ and G-Ref datasets.

In addition to comparing against full fine-tuning methods, we also evaluated our models in the context of parameter-efficient tuning methods. Table 1 shows that DETRIS-B and DETRIS-L both outperform existing parameter-efficient tuning methods such as ETRIS and BarLeRIa. Specifically, DETRIS-B achieves a substantial improvement, and DETRIS-L further enhances performance, demonstrating the effectiveness of our method.

Method	RefCOCO			Avg	Parameters (M)
	val	testA	testB		
Full-Tuning	65.1	68.1	61.4	64.9	149.97M
Fix Backbone	74.9	77.1	72.0	74.7	0.00 M
Adapter (Houlsby et al. 2019)	71.2	73.3	68.3	70.9	1.98M
Compacter (Karimi Mahabadi, Henderson, and Ruder 2021)	73.9	75.8	70.8	73.5	1.62M
LoRA (Hu et al. 2021)	73.4	75.7	70.2	73.1	1.57M
ETRIS (Xu et al. 2023)	74.5	76.5	72.9	74.6	1.38M
DETRIS-B (Ours)	75.8	77.7	72.9	75.5	1.36M
DETRIS-B (Ours) (Default Setting)	76.0	78.2	73.5	75.9	2.71M

Table 2: Comparison of Parameter-Efficient Tuning Methods Using DINO-B as Backbone on RefCOCO. To ensure fairness, we kept the original parameter settings from prior methods and also adjusted the size of rank to achieve comparable parameter counts.

We also evaluated our models on mixed RefCOCO datasets to test generalization. As shown in Table 1, DETRIS-L achieves the highest average IoU of 77.2, outperforming methods like PolyFormer-L and UNINEXT-L. This highlights the robustness and effectiveness of our approach, particularly as data volume increases. Our DETRIS models offer significant improvements in both IoU and parameter efficiency over existing RIS methods.

4.4 Qualitative Analysis

As shown in Figure 2, we present qualitative results with different settings across various scenarios. Panel (c) shows the baseline ETRIS method, which struggles with accurate localization and segmentation. In contrast, panel (d) highlights DETRIS-B with only the Text Adapter, achieving improved segmentation for text-driven descriptions, while panel (e) demonstrates DETRIS-B with only the Dense Aligner, providing better visual-text alignment. Panel (f), combining both adapters, achieves the most accurate and balanced results, while panel (g) displays DETRIS-L trained on mixed datasets, offering the best overall performance. Our methods significantly improve over ETRIS, especially in challenging scenarios, addressing limitations in object boundaries and ambiguous descriptions.

4.5 Ablation Study

Comparison with Other Parameter-Efficient Tuning Methods. We compare our Dense Aligner and Text Adapter (DETRIS-B) approach with other parameter-efficient tuning methods using DINO-B as the backbone. To ensure fairness, we retain the original parameter settings from prior methods and adjust rank size for comparable parameter counts. As shown in Table 2, DETRIS-B achieves superior performance while remaining efficient. Adapter reaches 70.9 with 1.98M parameters, Compacter 73.5 with 1.62M, LoRA 73.1 with 1.57M, and ETRIS 74.6 with 1.38M. Fully fine-tuning the misaligned DINO backbone degrades performance, underscoring the challenges of using DINO for RIS tasks. Our DETRIS-B surpasses these methods, achieving 75.5 with 1.36M parameters (0.9% of the backbone) by reducing Dense Aligners from 6 to 3. With 2.71M parameters (1.8% of the backbone), DETRIS-B achieves the highest score of 75.9. These results highlight the limitations of prior methods, which rely on pre-trained vision-text alignment and struggle in downstream RIS tasks with misaligned backbones. In contrast, our Dense Aligner actively learns cross-modal information during fine-tuning, overcoming these challenges effectively.

DA	TA	RefCOCO			Avg
		val	testA	testB	
×	×	74.9	77.1	71.9	74.6
✓	×	75.4	77.3	72.5	75.0
×	✓	75.9	77.5	72.7	75.4
✓	✓	76.0	78.2	73.5	75.9

Table 3: Ablation study on the components of DETRIS. DA stands for Dense Aligner, and TA denotes Text Adapter.

Effect of Dense Aligner and Text Adapter. We evaluated the Dense Aligner (DA) and Text Adapter (TA) through an ablation study on validation and test datasets. As shown in Table 3, the baseline model without adapters achieves the lowest average score of 74.6. Adding the Dense Aligner slightly improves performance to 75.0, highlighting its role in enhancing visual feature extraction. The Text Adapter alone yields a greater improvement, raising the score to 75.4, demonstrating its importance in processing textual information. Combining both adapters achieves the highest score of 75.9, highlighting their combined effectiveness in improving cross-modal feature learning and segmentation.

5 Conclusion

In this work, we introduce DETRIS, a parameter-efficient tuning framework for referring image segmentation (RIS). Our approach adapts the pre-trained DINO model for RIS, focusing on aligning encoders not pre-trained for multi-modal tasks. We propose the Dense Aligner to enhance fine-grained visual-text alignment and use text adapters to augment the language encoder. Our method outperforms state-of-the-art fully fine-tuned models while being more scalable and efficient in parameter management.

Acknowledgments

This work was supported by the Guangdong Provincial Natural Science Foundation Project under Grant 2022A1515011120.

References

Chen, S.; Ge, C.; Tong, Z.; Wang, J.; Song, Y.; Wang, J.; and Luo, P. 2022a. Adaptformer: Adapting vision transformers for scalable visual recognition. *arXiv preprint arXiv:2205.13535*.

Chen, X.; Wang, X.; Changpinyo, S.; Piergiovanni, A.; Padlewski, P.; Salz, D.; Goodman, S.; Grycner, A.; Mustafa, B.; Beyer, L.; et al. 2022b. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Chen, Z.; Duan, Y.; Wang, W.; He, J.; Lu, T.; Dai, J.; and Qiao, Y. 2022c. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*.

Chng, Y. X.; Zheng, H.; Han, Y.; Qiu, X.; and Huang, G. 2023. Mask Grounding for Referring Image Segmentation. *arXiv preprint arXiv:2312.12198*.

Ding, H.; Liu, C.; Wang, S.; and Jiang, X. 2022. VLT: Vision-Language Transformer and Query Generation for Referring Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Fang, C.; He, C.; Xiao, F.; Zhang, Y.; Tang, L.; Zhang, Y.; Li, K.; and Li, X. 2024. Real-world Image Dehazing with Coherence-based Label Generator and Cooperative Unfolding Network. *arXiv preprint arXiv:2406.07966*.

Fang, Y.; Wang, W.; Xie, B.; Sun, Q.; Wu, L.; Wang, X.; Huang, T.; Wang, X.; and Cao, Y. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19358–19369.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2021. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*.

Guo, D.; Rush, A. M.; and Kim, Y. 2020. Parameter-efficient transfer learning with diff pruning. *arXiv preprint arXiv:2012.07463*.

Hao, T.; Chen, H.; Guo, Y.; and Ding, G. 2023. Consolidator: Mergable Adapter with Group Connections for Vision Transformer. In *International Conference on Learning Representations*.

He, C.; Li, K.; Zhang, Y.; Tang, L.; Zhang, Y.; Guo, Z.; and Li, X. 2023. Camouflaged object detection with feature decomposition and edge reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22046–22055.

He, C.; Li, K.; Zhang, Y.; Xu, G.; Tang, L.; Zhang, Y.; Guo, Z.; and Li, X. 2024a. Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. *Advances in Neural Information Processing Systems*, 36.

He, C.; Shen, Y.; Fang, C.; Xiao, F.; Tang, L.; Zhang, Y.; Zuo, W.; Guo, Z.; and Li, X. 2024b. Diffusion Models in Low-Level Vision: A Survey. *arXiv preprint arXiv:2406.11138*.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, 2790–2799. PMLR.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hu, Y.; Wang, Q.; Shao, W.; Xie, E.; Li, Z.; Han, J.; and Luo, P. 2023. Beyond One-to-One: Rethinking the Referring Image Segmentation. arXiv:2308.13853.

Ji, L.; Du, Y.; Dang, Y.; Gao, W.; and Zhang, H. 2024. A survey of methods for addressing the challenges of referring image segmentation. *Neurocomputing*, 583: 127599.

Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; and Carion, N. 2021. MDETR-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1780–1790.

Karimi Mahabadi, R.; Henderson, J.; and Ruder, S. 2021. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems*, 34: 1022–1035.

Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 787–798.

Kim, N.; Kim, D.; Lan, C.; Zeng, W.; and Kwak, S. 2022. ReSTR: Convolution-free Referring Image Segmentation Using Transformers. In *CVPR*, 18145–18154.

Kim, S.; Kang, M.; Kim, D.; Park, J.; and Kwak, S. 2024. Extending CLIP's Image-Text Alignment to Referring Image Segmentation. arXiv:2306.08498.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2023. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*.

Li, H.; Cao, M.; Cheng, X.; Li, Y.; Zhu, Z.; and Zou, Y. 2023. G2L: Semantically Aligned and Uniform Video Grounding via Geodesic and Game Theory. In *International Conference on Computer Vision (ICCV), Oral*.

Li, H.; Cao, M.; Cheng, X.; Li, Y.; Zhu, Z.; and Zou, Y. 2024a. Exploiting auxiliary caption for video grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18508–18516.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 12888–12900. PMLR.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2022b. Grounded language-image pre-training. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10965–10975.

Li, X.; Qiu, K.; Wang, J.; Xu, X.; Singh, R.; Yamazak, K.; Chen, H.; Huang, X.; and Raj, B. 2024b. R2-Bench: Benchmarking the Robustness of Referring Perception Models under Perturbations. *arXiv preprint arXiv:2403.04924*.

Li, X. L.; and Liang, P. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Liu, C.; Ding, H.; and Jiang, X. 2023a. GRES: Generalized referring expression segmentation. In *CVPR*, 23592–23601.

Liu, C.; Ding, H.; and Jiang, X. 2023b. GRES: Generalized Referring Expression Segmentation. arXiv:2306.00968.

Liu, J.; Ding, H.; Cai, Z.; Zhang, Y.; Satzoda, R. K.; Mahadevan, V.; and Manmatha, R. 2023a. PolyFormer: Referring image segmentation as sequential polygon generation. In *CVPR*, 18653–18663.

Liu, T.; Liu, X.; Huang, S.; Chen, H.; Yin, Q.; Qin, L.; Wang, D.; and Hu, Y. 2024a. DARA: Domain-and Relation-aware Adapters Make Parameter-efficient Tuning for Visual Grounding. *arXiv preprint arXiv:2405.06217*.

Liu, T.; Liu, X.; Huang, S.; Shi, L.; Xu, Z.; Xin, Y.; Yin, Q.; and Liu, X. 2024b. Sparse-Tuning: Adapting vision transformers with efficient fine-tuning and inference. *arXiv preprint arXiv:2405.14700*.

Liu, T.; Xu, Z.; Hu, Y.; Shi, L.; Wang, Z.; and Yin, Q. 2024c. MaPPER: Multimodal Prior-guided Parameter Efficient Tuning for Referring Expression Comprehension. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 4984–4994.

Liu, Y.; Zhang, C.; Wang, Y.; Wang, J.; Yang, Y.; and Tang, Y. 2023b. Universal Segmentation at Arbitrary Granularity with Language Instruction. *arXiv preprint arXiv:2312.01623*.

Ma, Y.; Wang, Y.; Wu, Y.; Lyu, Z.; Chen, S.; Li, X.; and Qiao, Y. 2022a. Visual knowledge graph for human action reasoning in videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4132–4141.

Ma, Y.; Yang, T.; Shan, Y.; and Li, X. 2022b. Simvtp: Simple video text pre-training with masked autoencoders. *arXiv preprint arXiv:2212.03490*.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H. V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; HAZIZA, D.; Massa, F.; El-Nouby, A.; et al. 2023. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Sun, Q.; Wang, J.; Yu, Q.; Cui, Y.; Zhang, F.; Zhang, X.; and Wang, X. 2023. EVA-CLIP-18B: Scaling CLIP to 18 Billion Parameters. *arXiv preprint arXiv:2402.04252*.

Tang, J.; Zheng, G.; Shi, C.; and Yang, S. 2023. Contrastive grouping with transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23570–23580.

Wang, Y.; Li, J.; ZHANG, X.; Shi, B.; Li, C.; Dai, W.; Xiong, H.; and Tian, Q. 2023. BarLeRIa: An Efficient Tuning Framework for Referring Image Segmentation. In *The Twelfth International Conference on Learning Representations*.

Wang, Z.; Lu, Y.; Li, Q.; Tao, X.; Guo, Y.; Gong, M.; and Liu, T. 2021. CRIS: CLIP-Driven Referring Image Segmentation. *arXiv preprint arXiv:2111.15174*.

Wang, Z.; Lu, Y.; Li, Q.; Tao, X.; Guo, Y.; Gong, M.; and Liu, T. 2022. Cris: Clip-driven referring image segmentation. In *CVPR*, 11686–11695.

Wu, X.; Li, H.; Luo, Y.; Cheng, X.; Zhuang, X.; Cao, M.; and Fu, K. 2024. Uncertainty-aware sign language video retrieval with probability distribution modeling. *ECCV 2024*.

Xu, Z.; Chen, Z.; Zhang, Y.; Song, Y.; Wan, X.; and Li, G. 2023. Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17503–17512.

Yan, B.; Jiang, Y.; Wu, J.; Wang, D.; Luo, P.; Yuan, Z.; and Lu, H. 2023. Universal instance perception as object discovery and retrieval. In *CVPR*, 15325–15336.

Yang, Y.; Ma, C.; Yao, J.; Zhong, Z.; Zhang, Y.; and Wang, Y. 2024. Remamber: Referring image segmentation with mamba twister. *arXiv preprint arXiv:2403.17839*.

Yang, Z.; Wang, J.; Tang, Y.; Chen, K.; Zhao, H.; and Torr, P. H. 2022. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, 18155–18165.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision*, 69–85. Springer.

Zaken, E. B.; Ravfogel, S.; and Goldberg, Y. 2021. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*.

Zhao, W.; Rao, Y.; Liu, Z.; Liu, B.; Zhou, J.; and Lu, J. 2023. Unleashing Text-to-Image Diffusion Models for Visual Perception. *ICCV*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhuang, X.; Li, H.; Cheng, X.; Zhu, Z.; Xie, Y.; and Zou, Y. 2025. Kdpror: A knowledge-decoupling probabilistic framework for video-text retrieval. In *European Conference on Computer Vision*, 313–331. Springer.

Zou, X.; Yang, J.; Zhang, H.; Li, F.; Li, L.; Gao, J.; and Lee, Y. J. 2023. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*.