
SaFiRe: Saccade-Fixation Reiteration with Mamba for Referring Image Segmentation

Zhenjie Mao¹ Yuhuan Yang¹ Chaofan Ma¹ Dongsheng Jiang⁴
Jiangchao Yao^{1†} Ya Zhang^{2,3†} Yanfeng Wang²

¹Cooperative Medianet Innovation Center, Shanghai Jiao Tong University

²School of Artificial Intelligence, Shanghai Jiao Tong University

³Institute of Artificial Intelligence for Medicine, Shanghai Jiao Tong University School of Medicine

⁴Huawei Inc.

{zhenjmao, yangyuhuan, chaofanma, sunarker, ya_zhang, wangyanfeng622}@sjtu.edu.cn
dongsheng_jiang@outlook.com

Abstract

Referring Image Segmentation (RIS) aims to segment the target object in an image given a natural language expression. While recent methods leverage pre-trained vision backbones and more training corpus to achieve impressive results, they predominantly focus on simple expressions—short, clear noun phrases like “red car” or “left girl”. This simplification often reduces RIS to a key word/concept matching problem, limiting the model’s ability to handle referential ambiguity in expressions. In this work, we identify two challenging real-world scenarios: *object-distracting expressions*, which involve multiple entities with contextual cues, and *category-implicit expressions*, where the object class is not explicitly stated. To address the challenges, we propose a novel framework, **SaFiRe**, which mimics the human two-phase cognitive process—first forming a global understanding, then refining it through detail-oriented inspection. This is naturally supported by Mamba’s scan-then-update property, which aligns with our phased design and enables efficient multi-cycle refinement with linear complexity. We further introduce **aRefCOCO**, a new benchmark designed to evaluate RIS models under ambiguous referring expressions. Extensive experiments on both standard and proposed datasets demonstrate the superiority of SaFiRe over state-of-the-art baselines. Project page: <https://zhenjiemao.github.io/SaFiRe/>.

1 Introduction

Referring Image Segmentation (RIS), aiming to segment the target object in an image based on the textual description, is a fundamental task in multi-modal understanding. Most studies solve the RIS task by designing image-text alignment modules to bridge the modality gap [1–3]. With the success of image foundation models [4–6], recent works further improved the performance of RIS by leveraging the pre-trained image features [7–9] and introducing more training corpus [10–12].

Despite significant progress, current RIS methods primarily focus on a *narrow* scenario, where most of the referring expressions consist of a *single, simple noun phrase* that *clearly* identifies the target object, *e.g.*, “left girl” [13], “red car” [13], etc. We summarize this as the **simple expression** pattern. Under this paradigm, the RIS task is often over-simplified and modeled as a “key word/concept matching problem”. The model only needs to extract the key words or concepts from the given text, then matches them with the image patches. Most existing RIS methods adopt this simplified

[†]Corresponding authors.

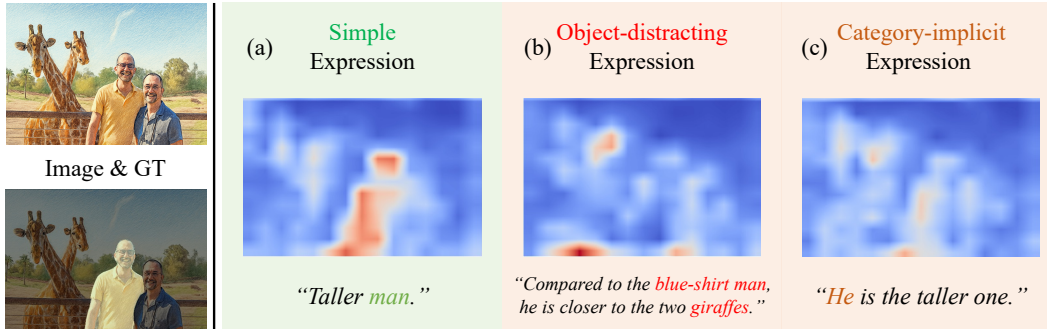


Figure 1: **Referential Ambiguity: One Object, Divergent Attention.** Attention maps under three types of referring expressions, all targeting the same object. (a) **Simple expression** yields accurate focus. (b) **Object-distracting expression** misguides attention to irrelevant regions. (c) **Category-implicit expression** leads to dispersed attention. This highlights the challenge of referential ambiguity for “key word/concept matching” method and motivates our saccade-fixation framework.

formulation. For example, LAVT [7] and ReLA [14] treat RIS as a process of matching individual words to specific pixels, matching keywords for each image location and segmenting the regions most relevant. VLT [15] introduces multiple queries to focus on various word pairs, and it reduces the rich semantics of language to a set of discrete choices. Similarly, MagNet [16] tries to emphasize key words by masking several input textual tokens and matching them with visual objects.

Although these methods demonstrate reasonable performance to some extent in this simple expression scenario, we notice that, in real-world applications, referring expressions often exhibit **referential ambiguity**, where the mapping between the text and the target object is indirect, underspecified, or contextually entangled. We summarize this into two challenging cases. (1) One is **object-distracting expression**, where the referring sentence contains multiple noun phrases but only one is the true referent, e.g., “compared to the blue-shirt man, he is closer to the two giraffes”. This introduces misleading contextual cues that can divert attention away from the actual target, increasing the difficulty of accurate segmentation. (2) The other is **category-implicit expression**, where the target object lacks an explicit categorical attribute, e.g., “he is the taller one”. This is common in spoken language, where pronouns and comparative or other distinct adjectives are frequently used without specifying object categories, making it difficult to determine the referent based solely on class-level semantics. As shown in Fig. 1, we visualize the attention maps under three types of expressions that refer to the same object. While the model successfully highlights the target under the simple expression (a), its focus becomes diffuse and uncertain under the object-distracting (b) and category-implicit (c) expressions. The referring expression in these cases is not ideally with a single object or a clear category, resulting in limited comprehension and weak visual-textual interaction for the existing methods that rely on key word/concept matching.

To address this issue, we investigate how humans interpret the correspondence between complex textual descriptions and visual representations, and how they identify objects within images. According to the research in cognitive psychology [17, 18], humans engage in a two-phase process. The first phase, which can be summarized as *global semantic understanding*, involves an initial overview of the sentence and a general scan of the image. The second phase, referred to as *cross-modal refinement*, occurs when the individual focuses on finer details, re-examining the sentence and systematically inspecting each region of the image to extract relevant information. This dual-phase process underscores how humans effectively integrate textual and visual information.

Inspired by this, we propose our network **SaFiRe** with two *alternating* operations: **Saccade** and **Fixation**, simulating the two human cognition phase respectively. This formulation aligns well with state-space sequence modeling, and we adopt Mamba as the underlying architecture to support our framework in a cognitively inspired and structurally efficient manner. Specifically, the **Saccade** operation performs a rapid “glimpse” of both the text and image. By modulating the image features with the general meaning of the text, this operation establishes a rough correspondence between the modalities after SSM scanning. While the **Fixation** operation aims to refine this coarse result by carefully examining the whole image region by region, and continuously reconfirms the text during this process. To this end, this operation alternates each image region with reiterated text in

a unified sequence for SSM scanning, enabling Mamba to focus locally while maintaining global context—naturally leveraging its order-sensitive and memory-efficient design. Furthermore, with the *reiteration* of the two-phase process, the model is able to progressively refine its cross-modal understanding and accurately identify the target referent.

To sum up, our contributions lie in four folds:

- We highlight the limitation of recent RIS methods, identifying their tendency to oversimplify RIS as a key word/concept matching problem, which often fails when confronted with referential ambiguity.
- We propose **SaFiRe**, a novel RIS framework inspired by cognitive psychology, which simulates the human cognitive process through alternating two inspection phases to achieve coarse-to-fine alignment and comprehensive cross-modal understanding.
- We design two operations—**Saccade** and **Fixation**—where Saccade establishes initial semantic correspondence through fast global scanning, and Fixation enables fine-grained region-wise inspection guided by reiterated textual cues for more accurate segmentation.
- To evaluate the effectiveness of our approach, we not only conduct experiments on the traditional RefCOCO family of datasets but also introduce a new test-only benchmark, **aRefCOCO**, which features ambiguous expressions. Extensive experiments and ablation studies demonstrate the superior performance of our method over the state-of-the-art methods.

2 Related Work

Referring Image Segmentation (RIS) is a challenging task that requires identifying and segmenting specific objects in images based on natural language expressions. Early methods mainly focus on attention-based vision-language fusion strategies. For example, VLT [15] introduced a cross-attention-based framework with query generation, while LAVT [7] demonstrated that encoder-level cross-modal fusion significantly enhances vision-language alignment. ReLA [14] advanced the field by incorporating fine-grained cross-region modality interactions. Later methods such as CGFormer [11] and MagNet [16] enhance segmentation accuracy by integrating object-level priors and mask supervision. The advent of Large Language Models (LLMs) has also spurred the development of LLM-based RIS models [19–21], leveraging their powerful multi-modal reasoning abilities to better address the RIS challenges. While existing methods perform well on simple expressions with clear object references, they often treat RIS as a key word/concept matching problem, which limits their ability to handle referential ambiguity.

State Space Models and Multi-Modal Extensions. Originally developed for dynamical systems, state-space models (SSMs) have proven effective in deep learning for modeling long-range dependencies. S4 [22] demonstrated their scalability, and Mamba [23] further improved efficiency and performance through a novel selection mechanism. Mamba has achieved strong results across vision [6, 24, 25] and language [26] tasks. Recently, its use in multi-modal learning [27–31] has gained interest, mainly in structurally aligned settings. However, current approaches offer limited cross-modal alignment. For instance, Coupled Mamba [32] uses shallow fusion strategies, while ReMamber [9] focuses on encoder-level fusion but oversimplifies local interaction as a keyword-matching problem. To address these gaps, we propose a framework that leverages Mamba’s sequential processing to emulate the human-like two-phase grounding process.

3 Method

3.1 Overview

The standard definition of RIS task is to segment the target object based on the input textual description. Given an image \mathbf{I} and a textual description \mathbf{T} , we aim to predict a binary mask \mathbf{M} that indicates the referring target in the image. The overall framework of our proposed can be written as:

$$\Phi_{\text{SaFiRe}}(\mathbf{I}, \mathbf{T}) = \mathbf{M}. \quad (1)$$

For encoders, we use Swin-Transformer [4] and BERT [33] to extract image and text features respectively,

$$\mathbf{F}_v^0 = \mathcal{E}_v(\mathbf{I}) \in \mathbb{R}^{H \times W \times C}, \quad \mathbf{F}_t^0 = \mathcal{E}_t(\mathbf{T}) \in \mathbb{R}^{L \times C}. \quad (2)$$

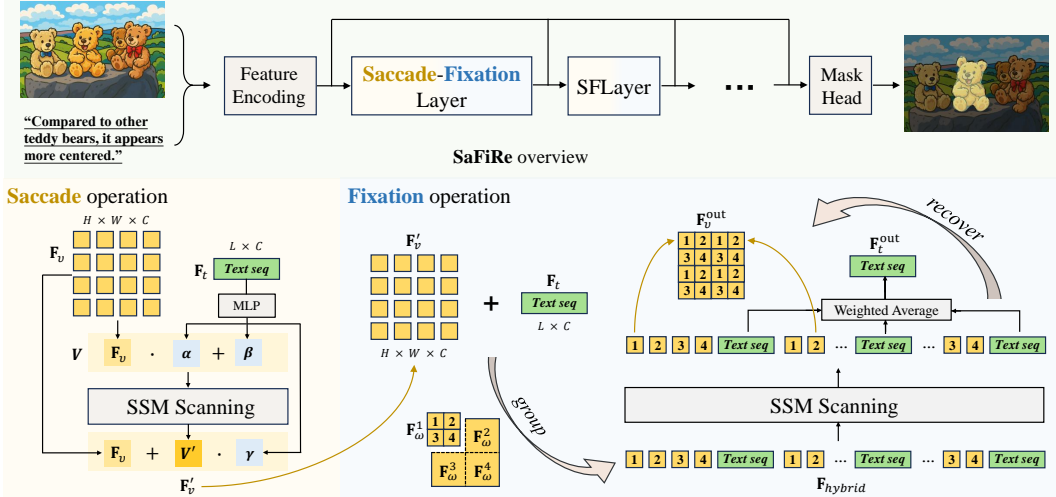


Figure 2: **Overview of the SaFiRe Architecture.** For each SFLayer, it consists of Saccade operation and Fixation operation. The **Saccade** operation corresponds to the phase of global semantic understanding. It enables the model to rapidly scan both visual and textual inputs, establishing a coarse-level alignment between the two modalities. The **Fixation** operation mirrors the cross-modal refinement phase. It allows the model to attend to specific local visual regions while re-examining the textual input, facilitating the extraction of fine-grained, task-relevant information.

Then, we decode the image and text features through a series of our carefully designed **SFLayers**:

$$\mathbf{F}_v^{i+1}, \mathbf{F}_t^{i+1} = \Phi_{\text{SFLayer}}^{(i)}(\mathbf{F}_v^i, \mathbf{F}_t^i), \quad i \in \{0, 1, 2, 3\}. \quad (3)$$

Finally we predict mask with processed image features:

$$\mathbf{M} = \Phi_{\text{mask}}(\mathbf{F}_v^1, \mathbf{F}_v^2, \mathbf{F}_v^3, \mathbf{F}_v^4). \quad (4)$$

SFLayer Design. Our **SaFiRe** framework simulates the two-phase human cognitive process by applying a specialized module called the *Saccade-Fixation Layer (SFLayer)*. Each SFLayer consists of two sequential operations: **Saccade** and **Fixation**. The **Saccade** operation corresponds to the phase of global semantic understanding. It enables the model to rapidly scan both visual and textual inputs, establishing a coarse-level alignment between the two modalities. In contrast, the **Fixation** operation mirrors the cross-modal refinement phase. It allows the model to attend to specific local visual regions while re-examining the textual input, facilitating the extraction of fine-grained, task-relevant information. Formally, Eq. (3) can be expanded in detail as:

$$\begin{aligned} \mathbf{F}_v^{i+1}, \mathbf{F}_t^{i+1} &= \Phi_{\text{SFLayer}}^{(i)}(\mathbf{F}_v^i, \mathbf{F}_t^i) \\ &= \Phi_{\text{Fixation}}(\Phi_{\text{Saccade}}(\mathbf{F}_v^i, \mathbf{F}_t^i)). \end{aligned} \quad (5)$$

By *reiteratively* alternating between these two operations across multiple layers, the model progressively enhances its multi-modal understanding. This saccade-fixation reiteration across layers serves as the core mechanism for accurate target localization. For a clearer view of this process, the architectural details of our framework are shown in Fig. 2, and we detail the design of **Saccade** and **Fixation** operations in Sec. 3.2 and Sec. 3.3 respectively.

3.2 Saccade: Quick Glimpse of Image-Text Correspondence

Our **Saccade** operation aims to take a *quick glimpse* of both text and image information, and establish a rough correspondence between them. This step is designed to mimic the initial phase of human referential understanding—namely, a *coarse and global semantic alignment* between modalities before any fine-grained reasoning occurs. We implement this by modulating image features with global textual semantics via a lightweight yet effective operation. Concretely, we adopt a DiT-inspired [34] Norm Adaptation, the pooled textual representation adjusts the statistics of image features, in contrast to class-based conditioning in the original design.

To be more specific, the Norm Adaptation operation pools a global vector from textual sequence and adapts the image’s scale and bias *globally* according to this vector. In our practice, with image feature $\mathbf{F}_v \in \mathbb{R}^{H \times W \times C}$ and text feature $\mathbf{F}_t \in \mathbb{R}^{L \times C}$, we apply average pooling and linear projection on text side to generate global scale, bias and shift factors:

$$[\alpha, \beta, \gamma] = \text{Linear}(\text{AvgPool}(\mathbf{F}_t)) \in \mathbb{R}^{3 \times C}. \quad (6)$$

The image feature is then adapted with the three factors:

$$\mathbf{F}'_v = \gamma \cdot \text{VSSM}(\alpha \cdot \mathbf{F}_v + \beta) + \mathbf{F}_v \in \mathbb{R}^{H \times W \times C}, \quad (7)$$

where $\text{VSSM}(\cdot)$ denotes the visual SSM feed-forward layer in VMamba [6].

This global adapting operation performs a quick and general interaction between the two modalities, making the image features more prepared for the subsequent **Fixation** operation.

3.3 Fixation: Detailed Examination of Cross-Modal Alignment

For **Fixation** operation, we aim to foster image-text communication in a more delicate manner, by checking the image region by region according to the text description to determine the target more accurately. We find that *the SSM mechanism is especially suitable for this*. The hidden states of SSM is updated by previous input and its scanning mechanism is a vivid simulation of human behavior: it focuses more on the recent input sequence while keeping long-term global memory in mind.

Thus, we utilize the most recent VSSM architecture VMamba [6], and came up with a novel **Fixation** operation on top of it. The Fixation operation follows a “**group-scan-recover**” pattern, where it first **(1) groups** the image sequence into non-overlapping local regions via window splitting, and then **(2) scans** each region and the text alternately with SSM, effectively modeling fine-grained cross-modal interactions within a unified multi-modal sequence, finally **(3) recovers** the processed unified multi-modal sequence back into separate image and text representations for the next operation. We here detail each of them in order.

Group. We aim to divide image sequence into several regions while maintaining their local structural priors. We draw inspiration from Swin [4] and use similar split mechanism with several non-overlapping windows. To be specific, with input image sequence $\mathbf{F}_v \in \mathbb{R}^{H \times W \times C}$ and window size w , we split the sequence into $P = HW/w^2$ windows:

$$\mathbf{F}_v = [\mathbf{F}_w^1, \mathbf{F}_w^2, \dots, \mathbf{F}_w^P], \quad \mathbf{F}_w^i \in \mathbb{R}^{w^2 \times C}. \quad (8)$$

As SSM itself offers global reception field, we do not apply sliding window mechanism as in Swin.

Scan. We aim to perform local and fine-grained scanning while foster cross-modal interaction. There is one key property under SSM mechanism: more recent tokens will be more relevant to the current hidden state than the previous ones. In other words, for SSM *the input order matters*. We make full utilization of this property and reiterate the text sequence multiple times within each image region.

To be specific, with input image feature $\mathbf{F}_v \in \mathbb{R}^{H \times W \times C}$ and text feature $\mathbf{F}_t \in \mathbb{R}^{L \times C}$, we re-arrange a hybrid multi-modal sequence $\mathbf{F}_{\text{hybrid}}$ in the following manner:

$$\mathbf{F}_{\text{hybrid}} = [\mathbf{F}_w^1, \mathbf{F}_t, \mathbf{F}_w^2, \mathbf{F}_t, \dots, \mathbf{F}_w^P, \mathbf{F}_t] \in \mathbb{R}^{(HW+PL) \times C}. \quad (9)$$

Then, we use an VSSM layer to process this hybrid sequence:

$$\mathbf{F}'_{\text{hybrid}} = \text{VSSM}(\mathbf{F}_{\text{hybrid}}) = [\mathbf{F}_w^{1'}, \mathbf{F}_t^{1'}, \mathbf{F}_w^{2'}, \mathbf{F}_t^{2'}, \dots, \mathbf{F}_w^{P'}, \mathbf{F}_t^{P'}]. \quad (10)$$

Due to the re-arrangement of hybrid sequence and the built-in SSM mechanism, the model will always be focused on the query target with the help of periodically reiterated text tokens. This encourages the model to focus on the description as a coherent and complete whole, preserving its continuity and full meaning, rather than discretizing it into isolated key words/concepts for global matching.

Recover. This separates the unified multi-modal sequence back into distinct image and text features, preparing them for the next layer. For image features we re-arrange them, and for text features, we apply weighted average across all repeated sentences as updated representation. The output feature can be expressed as:

$$\mathbf{F}_v^{\text{out}} = [\mathbf{F}_w^{1'}, \mathbf{F}_w^{2'}, \dots, \mathbf{F}_w^{P'}] \in \mathbb{R}^{H \times W \times C}, \quad \mathbf{F}_t^{\text{out}} = \sum_{o=1}^P w_o \mathbf{F}_t^{o'} / P \in \mathbb{R}^{L \times C}. \quad (11)$$

Here $\mathbf{w} = [w_1, \dots, w_P] \in \mathbb{R}^P$ is a learnable average weight. $\mathbf{F}_v^{\text{out}}$ and $\mathbf{F}_t^{\text{out}}$ are the output features and then passed to the next SFLayer for further processing.

The **Fixation** operation achieves the goal of delicate multi-modal understanding by fully utilizing the underlying mechanism of SSM. By repeating the text multiple times within the image sequence, the hidden state space is periodically refreshed, forcing the model to *continuously memorize and reinforce* its segmentation target. With the help of SSM mechanism, model is able to be more focused on *the current area* while ensures the *global receptive field* for segmentation.

Discussion: Why Mamba Fits Our Framework. Our framework focuses on simulating the sequential nature of human cognition process—alternating between global understanding and localized refinement. This requires a model that can processes information step-by-step. Mamba’s state-space scanning aligns perfectly with this paradigm, as it processes sequences in a directional, recurrent manner. Besides, its linear complexity also allows efficient, dense region-text interactions, supporting fine-grained reasoning without the overhead of attention. These advantages make Mamba a structurally appropriate and computationally efficient fit for our group-scan-recover Fixation operation and the overall Saccade-Fixation Reiteration framework.

3.4 Segmentation Head

For final mask prediction, we combine the multi-level visual output $[\mathbf{F}_v^1, \mathbf{F}_v^2, \mathbf{F}_v^3, \mathbf{F}_v^4]$ of each SFLayer in a top-down manner, following LAVT [7], as characterized by Eq.(4). The training process employs a combination of Dice loss [35] and Focal loss [36] as the composite loss function, as follows:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{Dice}} + (1 - \alpha) \cdot \mathcal{L}_{\text{Focal}}, \quad (12)$$

where the weighting coefficient α is set to 0.5 by default.

3.5 Complexity Analysis

Mamba-based architectures are known for their linear complexity in long sequences processing. However, since we introduce multiple times of text repeating in **Fixation** module, the overall sequence length will be slightly increased. Here we take a concise analysis of the computational complexity of our **Fixation** module. Formally, with image sequence length HW , text sequence length L and local window size w , the complexity is defined by the overall sequence length:

$$\text{Complexity} = O(HW + HW/w^2 \cdot L) = (1 + L/w^2) \cdot O(HW). \quad (13)$$

In practice, we set window size w to 4, and average text length L to around 15. The overall complexity is thus increased by a constant factor of 0.9, which is nearly negligible.

4 Experiments

4.1 Datasets and Metrics

Traditional Benchmarks. Following prior works [7, 16, 9], we systematically assess model performance on the widely used RefCOCO benchmarks: RefCOCO [13], RefCOCO+ [13] and RefCOCOg [37]. These datasets are all grounded in the MSCOCO [38] visual corpus but differ significantly in their linguistic properties. RefCOCO and RefCOCO+ feature concise referring expressions, averaging 3.6 words per phrase. Notably, RefCOCO+ omits absolute spatial terms (e.g., “left/right” or ordinal indicators), thereby introducing additional difficulty. In contrast, RefCOCOg features more descriptive and elaborate language, with expressions averaging 8.4 words and incorporating more references to locations and appearances, making it the most linguistically complex and challenging benchmark among the three.

aRefCOCO. We introduce aRefCOCO (ambiguous RefCOCO), a **test-only benchmark** constructed by reannotating a subset of images from the RefCOCO [13] and RefCOCOg [37] test splits with more challenging referring expressions. It is specifically designed to evaluate model performance on **object-distracting** and **category-implicit** scenarios—two aspects that remain difficult for existing models. The referring expressions in aRefCOCO are significantly more complex, averaging 12.6 words per sentence. To increase the difficulty of category-implicit references, the benchmark includes a higher frequency of pronouns (e.g., “it”, “he”, “she”), which obscure explicit category cues and demand deeper contextual understanding. To better support object-distracting evaluation, expressions exhibit more frequent use of prepositions (e.g., “and”, “to”, “of”), introducing richer relational structures

and more noun phrases that increase the likelihood of semantic confusion. Visually, aRefCOCO images contain 3.1 same-category distractors per image on average—87.5% more than RefCOCOg’s 1.6—requiring models to resolve denser visual ambiguities for accurate grounding. In terms of scale, aRefCOCO introduces 7,050 text-image pairs generated by *Qwen-VL-Max* [39] and meticulously filtered—a 40% expansion over RefCOCOg’s test split (5,023 pairs)—establishing a benchmark that challenges existing RIS models. For more details of aRefCOCO, see Appendix D.

Metrics. We adopt oIoU as the main metric and further incorporate mIoU and Precision@X ($X \in \{50, 70, 90\}$) for a more comprehensive evaluation, where Precision@X means the percentage of test images with an IoU score higher than X%.

4.2 Implementation Details

Our model employs the Swin Transformer-B [4] and BERT-B [33] as the vision and language encoders, respectively, following prior work [7, 11]. Swin-B is initialized with ImageNet pre-trained weights, and BERT-B uses the official checkpoints. For SFlayers, we adopt the official implementation of the VSSM from VMamba [6] and set the window size to 4×4 in the Fixation operation. The model is trained end-to-end for 50 epochs using the AdamW optimizer with a learning rate of $5e-5$ and weight decay of $1e-4$, along with a cosine learning rate scheduler. All experiments are conducted on 2 NVIDIA A100 GPUs.

4.3 Comparison with the State-of-the-Arts

Table 1: Comparison of RIS Methods Trained on Single Datasets across RefCOCO, RefCOCO+, and RefCOCOg with oIoU. U indicates UMD partition of RefCOCOg. The best performances are in bold.

Methods	Encoders		RefCOCOg (hard)		RefCOCO+ (medium)			RefCOCO (easy)		
	Visual	Textual	val(U)	test(U)	val	testA	testB	val	testA	testB
CRIS [40]	CLIP _{R101}	CLIP	56.56	57.38	57.94	64.05	48.42	67.35	71.54	62.16
VLT [15]	Swin-B	BERT	63.49	66.22	63.53	68.43	56.92	72.96	75.96	69.60
LAVT [7]	Swin-B	BERT	61.24	62.09	62.14	68.38	55.10	72.73	75.82	68.79
BKINet [41]	CLIP _{R101}	CLIP	64.21	63.77	64.91	69.88	53.39	73.22	76.43	69.42
ReLA [14]	Swin-B	BERT	65.00	65.97	66.04	71.02	57.65	73.92	76.48	70.18
SLViT [8]	SegNeXt-B	BERT	62.75	63.57	64.07	69.28	56.14	74.02	76.91	70.62
SADLR [42]	Swin-B	BERT	63.60	63.56	62.48	69.09	55.19	74.24	76.25	70.06
DMMI [43]	Swin-B	BERT	63.46	64.19	63.98	69.73	57.03	74.13	77.13	70.16
CGFormer [11]	Swin-B	BERT	64.68	64.09	64.54	71.00	57.14	74.75	77.30	70.64
RISCLIP [44]	CLIP-B	CLIP-B	64.10	65.09	65.53	70.61	55.49	73.57	76.46	69.76
MagNet [16]	Swin-B	BERT	65.36	66.03	66.16	71.32	58.14	75.24	78.24	71.05
LQMFormer [12]	Swin-B	BERT	64.73	66.04	65.91	71.84	57.59	74.16	76.82	71.04
ReMamber [9]	Mamba-B	CLIP	63.90	64.00	65.00	70.78	57.53	74.54	76.74	70.89
SaFiRe (Ours)	Swin-B	BERT	67.07	66.87	66.39	71.96	58.47	75.34	78.52	71.35

Traditional Benchmarks. Tab. 1 presents a comparison of our **SaFiRe** with state-of-the-art methods on the RefCOCO, RefCOCO+, and RefCOCOg datasets with oIoU metric, where all models are trained independently on single datasets. Results show that **SaFiRe** outperforms all other methods across all datasets. In particular, **SaFiRe** achieved a great improvement on RefCOCOg, where the task involves more complex and fine-grained textual descriptions compared to RefCOCO.

In addition, Tab. 2 compares **SaFiRe** with recent state-of-the-art methods trained under mixed-data configuration. The precise dataset mixtures differ across methods. To avoid conflating performance with training-corpus choices, we report each baseline under its official training configuration. In general, these models draw on subsets and combinations of widely used referring-expression and segmentation datasets—such as RefCOCO [13], RefCOCO+ [13], RefCOCOg [37], COCO [38], Object365 [54], ADE20K [55], COCO-Stuff [56], PACO-LVIS [57], PASCAL-Part [58], GranD [20], VOC2010 [59], MUSE [53], gRefCOCO [14], and COCO-Interactive [52]. Our training corpus follows a mixed-data setup based on the RefCOCO series and COCO-Stuff, which is comparatively smaller. However, the results show that **SaFiRe** achieves competitive results on all three benchmarks, particularly excelling on the more challenging RefCOCOg and RefCOCO+ benchmarks. Notably, it surpasses several SAM-enabled models (e.g., u-LLaVA, GSVA) across multiple splits.

aRefCOCO. We further conduct a *zero-shot transfer evaluation* of different models on the aRefCOCO. The expressions in aRefCOCO are notably ambiguous and more complex, frequently lacking explicit category indications and containing more distractors, which demands a sophisticated cross-modal

Table 2: **Comparison of Model Performance Trained on Mixed Datasets on Three Benchmarks.** U indicates UMD partition of RefCOCOg. The best performances are in **bold**. (ft) denotes models further finetuned on RefCOCO+/g after mix training. “Textual” and “Visual” refer to the textual encoder and the visual backbone, respectively. “LLM” denotes the large language model used as the textual reasoner.

Methods	Textual / LLM	Visual	RefCOCOg (hard)		RefCOCO+ (medium)			RefCOCO (easy)			
			val(U)	test(U)	val	testA	testB	val	testA	testB	
mIoU	EEVG [45]	BERT	Swin-B	71.5	71.9	71.4	75.6	64.6	77.5	79.6	75.3
	PromptRIS [46]	CLIP	SAM	69.2	70.5	71.1	76.6	64.3	78.1	81.2	74.6
	OneRef-B (ft) [47]	BEiT3-B	BEiT3-B	74.1	74.9	74.7	77.9	69.6	79.8	81.9	77.0
	C3VG [48]	BEiT3-B	BEiT3-B	76.3	77.1	77.1	79.6	72.4	81.4	82.9	79.1
	SaFiRe (Ours)	BERT	Swin-B	77.3	78.1	77.4	80.5	72.9	82.1	83.7	80.1
oIoU	MagNet [16]	BERT	Swin-B	67.8	69.3	68.1	73.6	61.8	76.6	78.3	72.2
	PolyFormer-L [49]	BERT	Swin-L	69.2	70.2	69.3	74.6	61.9	76.0	78.3	73.3
	UNINEXT-L [50]	BERT	ConvNeXt-L	73.4	73.7	70.0	74.9	62.6	80.3	82.6	77.8
	LISA (ft) [19]	LLaVA-7B	SAM	67.9	70.6	65.1	70.8	58.1	74.9	79.1	72.3
	GLaMM [20]	Vicuna-7B	SAM	74.2	74.9	72.6	78.7	64.6	79.5	83.2	76.9
	u-LLaVA [51]	Vicuna-7B	SAM	74.8	75.6	72.2	76.6	66.8	80.4	82.7	77.8
	PSALM [*] [52]	Phi-1.5	Swin-B	71.0	72.3	68.1	70.7	64.4	78.0	78.1	76.6
	GSVA (ft) [21]	Vicuna-13B	SAM	74.2	75.6	67.4	71.5	60.9	78.2	80.4	74.2
	PixelLM [53]	Llama2-13B	CLIP _{vis}	69.3	70.5	66.3	71.7	58.3	73.0	76.5	68.2
	SaFiRe (Ours)	BERT	Swin-B	75.8	76.8	74.8	78.9	68.3	81.4	83.6	78.2

* PSALM results are re-evaluated to align with standard settings [19], which predicts masks for each image-text pair for evaluation.

Table 3: **Performance Comparison on aRefCOCO.**

Methods	aRefCOCO				
	oIoU	mIoU	Pr@50	Pr@70	Pr@90
LAVT [7]	31.7	34.9	30.8	19.9	5.5
CGFormer [11]	54.0	60.3	67.0	59.7	27.2
ReMamber [9]	49.4	57.8	64.2	57.9	27.9
MagNet [16]	54.4	60.7	67.7	61.4	29.9
EEVG [45]	50.7	59.4	66.9	62.4	29.4
C3VG [48]	62.9	69.1	78.4	72.8	32.8
LISA(7B) [19]	49.0	53.8	58.7	49.2	20.6
LISA++(7B) [60]	45.3	50.4	51.5	43.5	19.7
PSALM [52]	60.1	67.3	75.7	71.8	38.7
SaFiRe (Ours)	65.1	71.4	78.9	73.6	45.2

Table 4: **Performance Comparison of Different Configurations.**

Configuration		RefCOCOg		RefCOCO	
Saccade	Fixation	oIoU	mIoU	oIoU	mIoU
✓	✓	67.07	69.68	75.34	77.02
✓		65.72	68.24	74.27	76.12
	✓	66.32	68.73	73.60	75.64
		64.48	67.72	72.92	75.17

understanding for precise segmentation. For fair comparison, all models have NOT been trained on aRefCOCO. As shown in Tab. 3, our model achieves the best overall performance on aRefCOCO, demonstrating its strong ability to precisely localize referred objects under ambiguous language conditions.

4.4 Ablation Study

Module Ablation. Tab. 4 summarizes the performance of different module configurations within our model. The results show that our full model achieves the highest oIoU scores on both RefCOCOg (descriptive sentences) and RefCOCO (concise phrases). Compared to independent module usage, the synergistic combination of Saccade and Fixation yields significant gains, highlighting their complementary effects. Removing both modules reduces the model to a naive baseline that simply concatenates textual and visual sequences for VSSM scanning, which performs worst among all variants, underscoring the essential role of both Saccade and Fixation operations.

Fixation Window Size. In Fixation operation, the approach is related to sequence arrangement for SSM scanning, where the ordering of input tokens impacts performance. With I representing individual image tokens and T representing the full sentence feature tokens, we conduct an ablation

Table 5: **Performance Comparison of Different Fixation Window Sizes.**

Configuration	RefCOCOg
Vanilla [I . . . I T]	64.38
Repeat [I . . . I T . . . T]	65.50
Fixate 2 × 2	66.49
Fixate 4 × 4 (ours)	67.07
Fixate 8 × 8	66.20

Table 6: **Performance Comparison of Different Backbones.**

Backbone	Model	RefCOCOg
ViT-B	baseline	55.55
ViT-B	SaFiRe	59.33 (+3.78)
Swin-B	SaFiRe	67.07
Swin-L	SaFiRe	68.16 (+1.09)

Table 7: **Comparison of Inference Efficiency.**

Models	GFLOPs	FPS(image/s)
CGFormer [11]	631	2.78
PSALM [52]	455	2.19
LISA(7B) [19]	4880	2.67
LISA++(7B) [60]	4943	1.29
SaFiRe (Ours)	384	8.26

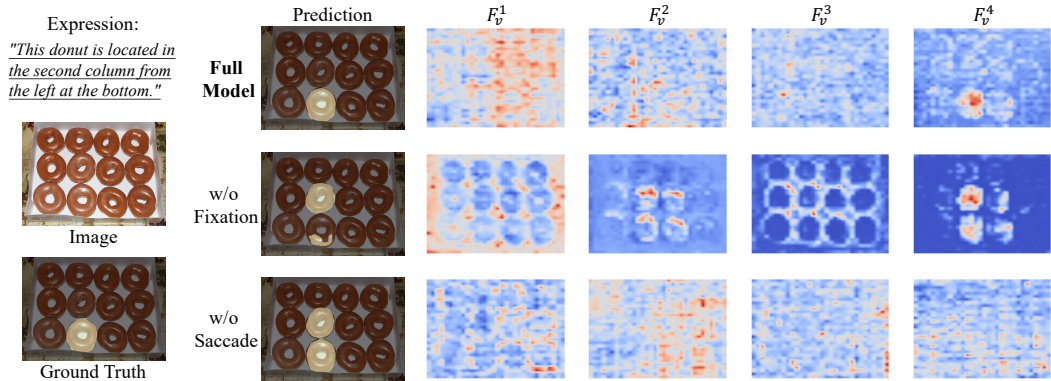


Figure 3: **Layer-Wise Visual Feature Maps.** Left→right corresponds to shallow→deep layers. The full model shows balanced activation. Without Fixation, local detail is missing; without Saccade, global focus weakens. The differing activation patterns reflect their complementary roles.



Figure 4: **Visualization Results for aRefCOCO.** Compared to the other two methods, our SaFiRe is more capable of comprehending ambiguous referring descriptions.

study of the following configurations in Tab. 5: (1) **Vanilla** [I . . . I T]: This configuration directly concatenates the image sequence [I . . . I] and the text sequence T. (2) **Repeat** [I . . . I T . . . T]: In this configuration, the text sequence is repeated multiple times at the end of the image sequence. (3) **Fixation** $x \times x$: This approach reiterates the text sequence once after every x^2 image tokens in window. For example, Fixation 2×2 means the sequence looks like: [IIII T IIII T . . .]. Tab. 5 shows that the Repeat configuration improves upon Vanilla, but the Fixation operation, particularly with a window size of 4×4 , shows a more significant enhancement in performance. This indicates that distributing textual information throughout the visual sequence helps maintain semantic continuity and guides the model to better align specific image regions with the text, rather than treating the two modalities as isolated blocks with simple concatenation.

Backbone Ablation. Here we present a comparison of different backbones, as depicted in Table Tab. 6. While our method is not specifically designed for global attention backbones such as ViT-B, it still brings a notable improvement of +3.78. Furthermore, when applied to stronger window-based backbones like Swin-L, the performance continues to improve, reaching 68.16 on RefCOCOg. These results indicate that our method generalizes well across different backbone architectures and particularly benefits from more powerful feature extractors.

Layer-wise Feature Map Analysis. In Fig. 3, we visualize the output visual feature maps of each SaFiRe layer under different model configurations. A clear difference in **activation patterns** can be observed between the Saccade and Fixation operations. Specifically, the Saccade operation generates broader and more globally distributed activations that align with the overall shape and location described in the text. In contrast, the Fixation operation produces sharper, more localized activations, focusing on fine-grained visual details. When used independently, each module exhibits limitations in fully capturing and grounding complex language expressions. However, when combined, their complementary activation behaviors allow the model to achieve an accurate identification of the

described object. This further confirms that Saccade and Fixation work in synergy to enhance the model’s cross-modal understanding.

4.5 Inference Efficiency Comparison

We evaluate inference efficiency under identical input settings and report both computational cost (GFLOPs) and throughput (FPS). As shown in Tab. 7, **SaFiRe** is compared against representative transformer-based (CGFormer [11]) and LLM-driven models (PSALM [52], LISA [19], LISA++ [60]). **SaFiRe** attains 8.26 FPS with 384 GFLOPs, yielding the highest throughput and the lowest compute among all compared methods. Together with the accuracy results in Tab. 1, Tab. 2 and Tab. 3, these findings underscore the efficiency advantage of our framework.

4.6 Visualization Results

Fig. 4 compares **SaFiRe** with two state-of-the-art RIS approaches—Mamba-based ReMamber [9] and LLM-driven PSALM [52]—on the **aRefCOCO** benchmark. Our method exhibits a stronger ability to handle referential ambiguity and avoids the confusion observed in the other two methods. For additional examples, see Appendix E.

4.7 Failure Case Analysis

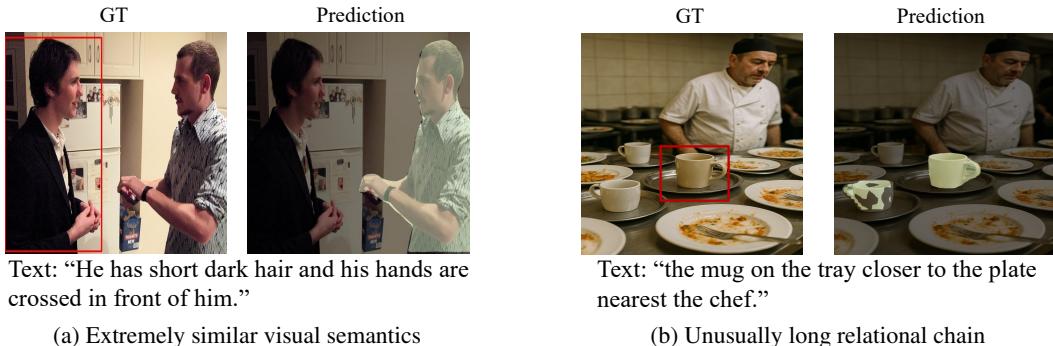


Figure 5: **Two Typical Types of Failure Cases.**

To guide future works, we analyze two typical failure scenarios. (1) Extremely similar visual semantics. Errors arise when multiple candidates share similar visual semantics but only one best fits the intended meaning. As in Fig. 5a, both people look “cross-like” at first glance; zooming in reveals that one is actually holding an object rather than truly crossing his hands. These cases demand particularly fine-grained visual discrimination and action-semantic grounding, which the model overlooked. (2) Unusually long relational chain. Failures occur in texts with very long spatial or relational chains. These require multi-step reasoning with frequent spatial shifts, where errors in intermediate links accumulate, as shown in Fig. 5b.

5 Conclusion

We present **SaFiRe**, a cognitively inspired framework for referring image segmentation that addresses the limitations of existing keyword-matching approaches. Motivated by the challenge of referential ambiguity, our method simulates the human two-phase grounding process through alternating *Saccade* and *Fixation* operations. Built upon the Mamba state-space model, **SaFiRe** enables efficient global-to-local semantic alignment and achieves state-of-the-art performance on both standard and ambiguous datasets, including the newly proposed aRefCOCO.

Limitation and Future Work. Despite its effectiveness, **SaFiRe** still faces challenges in handling language diversity in real-world application, which cannot be fully represented by existing datasets. Future work will explore incorporating structured scene understanding and pragmatic reasoning, and extending the framework to dialog-based and video grounding tasks. Moreover, **SaFiRe** has the potential to be adapted to other tasks or modalities with minimal modification, such as Open-Vocabulary Segmentation (OVS) [61–63] by adjusting the task head, or Audio-Visual Segmentation (AVS) [64–66] by substituting the backbone.

Acknowledgments

This work is supported by the National Key R&D Program of China (No. 2022ZD0160702), National Natural Science Foundation of China (No. 62306178) and STCSM (No. 22DZ2229005), 111 plan (No. BP0719010).

References

- [1] L. Ye, M. Rochan, Z. Liu, and Y. Wang, “Cross-modal self-attention network for referring image segmentation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 502–10 511.
- [2] G. Luo, Y. Zhou, R. Ji, X. Sun, J. Su, C.-W. Lin, and Q. Tian, “Cascade grouped attention network for referring expression segmentation,” *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.
- [3] Z. Hu, G. Feng, J. Sun, L. Zhang, and H. Lu, “Bi-directional relationship inferring network for referring image segmentation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4424–4433.
- [4] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.
- [5] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, “Segnext: Rethinking convolutional attention design for semantic segmentation,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 1140–1156, 2022.
- [6] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu, “Vmamba: Visual state space model,” *arXiv preprint arXiv:2401.10166*, 2024.
- [7] Z. Yang, J. Wang, Y. Tang, K. Chen, H. Zhao, and P. H. S. Torr, “LAVT: Language-Aware Vision Transformer for Referring Image Segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [8] S. Ouyang, H. Wang, S. Xie, Z. Niu, R. Tong, Y.-W. Chen, and L. Lin, “SLViT: Scale-Wise Language-Guided Vision Transformer for Referring Image Segmentation,” in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2023.
- [9] Y. Yang, C. Ma, J. Yao, Z. Zhong, Y. Zhang, and Y. Wang, “Remember: Referring image segmentation with mamba twister,” in *European Conference on Computer Vision*. Springer, 2025, pp. 108–126.
- [10] B. Chen, Z. Hu, Z. Ji, J. Bai, and W. Zuo, “Position-aware contrastive alignment for referring image segmentation,” *arXiv preprint arXiv:2212.13419*, 2022.
- [11] J. Tang, G. Zheng, C. Shi, and S. Yang, “Contrastive grouping with transformer for referring image segmentation,” in *CVPR*, 2023.
- [12] N. A. Shah, V. VS, and V. M. Patel, “Lqmformer: Language-aware query mask transformer for referring image segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12 903–12 913.
- [13] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, “ReferItGame: Referring to objects in photographs of natural scenes,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.
- [14] C. Liu, H. Ding, and X. Jiang, “Gres: Generalized referring expression segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 592–23 601.
- [15] H. Ding, C. Liu, S. Wang, and X. Jiang, “VLT: Vision-Language Transformer and Query Generation for Referring Segmentation,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 45, 2022.
- [16] Y. X. Chng, H. Zheng, Y. Han, X. Qiu, and G. Huang, “Mask grounding for referring image segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 573–26 583.

- [17] A. M. Treisman and G. Gelade, “A feature-integration theory of attention,” *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.
- [18] H. H. Clark and W. G. Chase, “On the process of comparing sentences against pictures,” *Cognitive Psychology*, vol. 3, no. 3, pp. 472–517, 1972.
- [19] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia, “Lisa: Reasoning segmentation via large language model,” *arXiv preprint arXiv:2308.00692*, 2023.
- [20] H. Rasheed, M. Maaz, S. Shaji, A. Shaker, S. Khan, H. Cholakkal, R. M. Anwer, E. Xing, M.-H. Yang, and F. S. Khan, “Gamm: Pixel grounding large multimodal model,” *arXiv preprint arXiv:2311.03356*, 2023.
- [21] Z. Xia, D. Han, Y. Han, X. Pan, S. Song, and G. Huang, “Gsva: Generalized segmentation via multimodal large language models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3858–3869.
- [22] A. Gu, K. Goel, and C. Ré, “Efficiently modeling long sequences with structured state spaces,” *arXiv preprint arXiv:2111.00396*, 2021.
- [23] A. Gu and T. Dao, “Mamba: Linear-time sequence modeling with selective state spaces,” *arXiv preprint arXiv:2312.00752*, 2023.
- [24] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, “Vision mamba: Efficient visual representation learning with bidirectional state space model,” *arXiv preprint arXiv:2401.09417*, 2024.
- [25] H. Guo, J. Li, T. Dai, Z. Ouyang, X. Ren, and S.-T. Xia, “Mambair: A simple baseline for image restoration with state-space model,” in *European Conference on Computer Vision*. Springer, 2025, pp. 222–241.
- [26] J. Park, J. Park, Z. Xiong, N. Lee, J. Cho, S. Oymak, K. Lee, and D. Papailiopoulos, “Can mamba learn how to learn? a comparative study on in-context learning tasks,” *arXiv preprint arXiv:2402.04248*, 2024.
- [27] S. Peng, X. Zhu, H. Deng, L.-J. Deng, and Z. Lei, “Fusionmamba: Efficient remote sensing image fusion with state space model,” *IEEE Transactions on Geoscience and Remote Sensing*, 2024.
- [28] X. Xie, Y. Cui, C.-I. Jeong, T. Tan, X. Zhang, X. Zheng, and Z. Yu, “Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba,” *arXiv preprint arXiv:2404.09498*, 2024.
- [29] W. Dong, H. Zhu, S. Lin, X. Luo, Y. Shen, X. Liu, J. Zhang, G. Guo, and B. Zhang, “Fusion-mamba for cross-modality object detection,” *arXiv preprint arXiv:2404.09146*, 2024.
- [30] H. Li, Q. Hu, Y. Yao, K. Yang, and P. Chen, “Cfmw: Cross-modality fusion mamba for multispectral object detection under adverse weather conditions,” *arXiv preprint arXiv:2404.16302*, 2024.
- [31] Z. Li, H. Pan, K. Zhang, Y. Wang, and F. Yu, “Mambadfuse: A mamba-based dual-phase model for multi-modality image fusion,” *arXiv preprint arXiv:2404.08406*, 2024.
- [32] W. Li, H. Zhou, J. Yu, Z. Song, and W. Yang, “Coupled mamba: Enhanced multi-modal fusion with coupled state space model,” *arXiv preprint arXiv:2405.18014*, 2024.
- [33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [34] W. Peebles and S. Xie, “Scalable diffusion models with transformers,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4195–4205.
- [35] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.
- [36] T.-Y. Ross and G. Dollár, “Focal loss for dense object detection,” in *proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2980–2988.
- [37] V. K. Nagaraja, V. I. Morariu, and L. S. Davis, “Modeling context between objects for referring expression understanding,” in *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 2016, pp. 792–807.
- [38] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

- [39] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, “Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond,” *arXiv preprint arXiv:2308.12966*, vol. 1, no. 2, p. 3, 2023.
- [40] Z. Wang, Y. Lu, Q. Li, X. Tao, Y. Guo, M. Gong, and T. Liu, “CRIS: CLIP-Driven Referring Image Segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [41] H. Ding, S. Zhang, Q. Wu, S. Yu, J. Hu, L. Cao, and R. Ji, “Bilateral Knowledge Interaction Network for Referring Image Segmentation,” *IEEE Transactions on Multimedia (TMM)*, vol. 26, 2023.
- [42] Z. Yang, J. Wang, Y. Tang, K. Chen, H. Zhao, and P. H. S. Torr, “Semantics-Aware Dynamic Localization and Refinement for Referring Image Segmentation,” in *AAAI Conference on Artificial Intelligence (AAAI)*, 2023.
- [43] Y. Hu, Q. Wang, W. Shao, E. Xie, Z. Li, J. Han, and P. Luo, “Beyond One-to-One: Rethinking the Referring Image Segmentation,” in *Proceedings of International Conference on Computer Vision (ICCV)*, 2023.
- [44] S. Kim, M. Kang, D. Kim, J. Park, and S. Kwak, “Extending clip’s image-text alignment to referring image segmentation,” *arXiv preprint arXiv:2306.08498*, 2024.
- [45] W. Chen, L. Chen, and Y. Wu, “An efficient and effective transformer decoder-based framework for multi-task visual grounding,” in *European Conference on Computer Vision*. Springer, 2024, pp. 125–141.
- [46] C. Shang, Z. Song, H. Qiu, L. Wang, F. Meng, and H. Li, “Prompt-driven referring image segmentation with instance contrasting,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4124–4134.
- [47] L. Xiao, X. Yang, F. Peng, Y. Wang, and C. Xu, “Oneref: Unified one-tower expression grounding and segmentation with mask referring modeling,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 139 854–139 885, 2024.
- [48] M. Dai, J. Li, J. Zhuang, X. Zhang, and W. Yang, “Multi-task visual grounding with coarse-to-fine consistency constraints,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 3, 2025, pp. 2618–2626.
- [49] J. Liu, H. Ding, Z. Cai, Y. Zhang, R. K. Satzoda, V. Mahadevan, and R. Manmatha, “Polyformer: Referring image segmentation as sequential polygon generation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 18 653–18 663.
- [50] B. Yan, Y. Jiang, J. Wu, D. Wang, P. Luo, Z. Yuan, and H. Lu, “Universal instance perception as object discovery and retrieval,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 325–15 336.
- [51] J. Xu, L. Xu, Y. Yang, X. Li, Y. Xie, Y.-J. Huang, and Y. Li, “u-llava: Unifying multi-modal tasks via large language model,” *arXiv preprint arXiv:2311.05348*, 2023.
- [52] Z. Zhang, Y. Ma, E. Zhang, and X. Bai, “Psalm: Pixelwise segmentation with large multi-modal model,” in *European Conference on Computer Vision*. Springer, 2025, pp. 74–91.
- [53] Z. Ren, Z. Huang, Y. Wei, Y. Zhao, D. Fu, J. Feng, and X. Jin, “Pixellm: Pixel reasoning with large multimodal model,” *arXiv preprint arXiv:2312.02228*, 2023.
- [54] S. Shao, Z. Li, T. Zhang, C. Peng, G. Yu, X. Zhang, J. Li, and J. Sun, “Objects365: A large-scale, high-quality dataset for object detection,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8430–8439.
- [55] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, “Semantic understanding of scenes through the ade20k dataset,” *International Journal of Computer Vision*, vol. 127, pp. 302–321, 2019.
- [56] H. Caesar, J. Uijlings, and V. Ferrari, “Coco-stuff: Thing and stuff classes in context,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1209–1218.
- [57] V. Ramanathan, A. Kalia, V. Petrovic, Y. Wen, B. Zheng, B. Guo, R. Wang, A. Marquez, R. Kovvuri, A. Kadian *et al.*, “Paco: Parts and attributes of common objects,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7141–7151.
- [58] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille, “Detect what you can: Detecting and representing objects using holistic models and body parts,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1971–1978.

- [59] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results,” <http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html>.
- [60] S. Yang, T. Qu, X. Lai, Z. Tian, B. Peng, S. Liu, and J. Jia, “An improved baseline for reasoning segmentation with large language model,” *CoRR*, 2023.
- [61] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, “Language-driven semantic segmentation,” *arXiv preprint arXiv:2201.03546*, 2022.
- [62] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin, “Scaling open-vocabulary image segmentation with image-level labels,” in *European conference on computer vision*. Springer, 2022, pp. 540–557.
- [63] C. Ma, Y. Yang, Y. Wang, Y. Zhang, and W. Xie, “Open-vocabulary semantic segmentation with frozen vision-language models,” *arXiv preprint arXiv:2210.15138*, 2022.
- [64] J. Zhou, J. Wang, J. Zhang, W. Sun, J. Zhang, S. Birchfield, D. Guo, L. Kong, M. Wang, and Y. Zhong, “Audio–visual segmentation,” in *European Conference on Computer Vision*. Springer, 2022, pp. 386–403.
- [65] J. Liu, Y. Wang, C. Ju, C. Ma, Y. Zhang, and W. Xie, “Annotation-free audio-visual segmentation,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 5604–5614.
- [66] J. Liu, C. Ju, C. Ma, Y. Wang, Y. Wang, and Y. Zhang, “Audio-aware query-enhanced transformer for audio-visual segmentation,” *arXiv preprint arXiv:2307.13236*, 2023.
- [67] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
- [68] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [69] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.
- [70] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.
- [71] C. Ma, Q. Xu, X. Wang, B. Jin, X. Zhang, Y. Wang, and Y. Zhang, “Boundary-aware supervoxel-level iteratively refined interactive 3d image segmentation with multi-agent reinforcement learning,” *IEEE Transactions on Medical Imaging*, vol. 40, no. 10, pp. 2563–2574, 2020.
- [72] Z. Ye, C. Ju, C. Ma, and X. Zhang, “Unsupervised domain adaption via similarity-based prototypes for cross-modality segmentation,” in *MICCAI Workshop on Domain Adaptation and Representation Transfer*. Springer, 2021, pp. 133–143.
- [73] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, “Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.
- [74] R. Ranftl, A. Bochkovskiy, and V. Koltun, “Vision transformers for dense prediction,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 179–12 188.
- [75] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, “Segformer: Simple and efficient design for semantic segmentation with transformers,” *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.
- [76] B. Cheng, A. Schwing, and A. Kirillov, “Per-pixel classification is not all you need for semantic segmentation,” *Advances in neural information processing systems*, vol. 34, pp. 17 864–17 875, 2021.
- [77] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, “Segment anything,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.
- [78] W. Liu, C. Ma, Y. Yang, W. Xie, and Y. Zhang, “Transforming the interactive segmentation for medical imaging,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 704–713.

- [79] D. Baranchuk, I. Rubachev, A. Voynov, V. Khruikov, and A. Babenko, “Label-efficient semantic segmentation with diffusion models,” *arXiv preprint arXiv:2112.03126*, 2021.
- [80] J. Wolleb, R. Sandkühler, F. Bieder, P. Valmaggia, and P. C. Cattin, “Diffusion models for implicit image segmentation ensembles,” in *International conference on medical imaging with deep learning*. PMLR, 2022, pp. 1336–1348.
- [81] T. Amit, T. Shaharbany, E. Nachmani, and L. Wolf, “Segdiff: Image segmentation with diffusion probabilistic models,” *arXiv preprint arXiv:2112.00390*, 2021.
- [82] J. Wu, R. Fu, H. Fang, Y. Zhang, Y. Yang, H. Xiong, H. Liu, and Y. Xu, “Medsegdiff: Medical image segmentation with diffusion probabilistic model,” in *Medical Imaging with Deep Learning*. PMLR, 2024, pp. 1623–1639.
- [83] T. Zhang, F. Zhang, J. Yao, Y. Zhang, and Y. Wang, “G4seg: Generation for inexact segmentation refinement with diffusion models,” *arXiv preprint arXiv:2506.01539*, 2025.
- [84] C. Ma, Y. Yang, C. Ju, Y. Shi, Y. Zhang, and Y. Wang, “Freeseqdiff: Annotation-free saliency segmentation with diffusion models,” in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
- [85] C. Ma, Y. Yang, C. Ju, F. Zhang, J. Liu, Y. Wang, Y. Zhang, and Y. Wang, “Diffusionseg: Adapting diffusion towards unsupervised object discovery,” *arXiv preprint arXiv:2303.09813*, 2023.
- [86] T. Zhang, C. Ma, and Y. Wang, “Tracking the rareness of diseases: Improving long-tail medical detection with a calibrated diffusion model,” *Electronics*, vol. 13, no. 23, p. 4693, 2024.
- [87] C. Ma, Y. Yang, C. Ju, F. Zhang, Y. Zhang, and Y. Wang, “Attrseg: open-vocabulary semantic segmentation via attribute decomposition-aggregation,” *Advances in neural information processing systems*, vol. 36, pp. 10 258–10 270, 2023.
- [88] J. Xu, S. De Mello, S. Liu, W. Byeon, T. Breuel, J. Kautz, and X. Wang, “Groupvit: Semantic segmentation emerges from text supervision,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 134–18 144.
- [89] F. Zhang, T. Zhou, B. Li, H. He, C. Ma, T. Zhang, J. Yao, Y. Zhang, and Y. Wang, “Uncovering prototypical knowledge for weakly open-vocabulary semantic segmentation,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 73 652–73 665, 2023.
- [90] Y. Yang, C. Ma, C. Ju, F. Zhang, J. Yao, Y. Zhang, and Y. Wang, “Multi-modal prototypes for open-world semantic segmentation,” *International Journal of Computer Vision*, vol. 132, pp. 6004–6020, 2024.
- [91] L. Liu, M. Zhang, J. Yin, T. Liu, W. Ji, Y. Piao, and H. Lu, “Defmamba: Deformable visual state space model,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 8838–8847.
- [92] J. Wang, D. Paliotta, A. May, A. Rush, and T. Dao, “The mamba in the llama: Distilling and accelerating hybrid models,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 62 432–62 457, 2024.
- [93] L. Wang, B. Zhu, Y. Chen, Y. Zhang, M. Tang, and J. Wang, “Mug: Pseudo labeling augmented audio-visual mamba network for audio-visual video parsing,” *arXiv preprint arXiv:2507.01384*, 2025.
- [94] Y. Yang, C. Ma, Z. Mao, J. Yao, Y. Zhang, and Y. Wang, “Moma: Modulating mamba for adapting image foundation models to video recognition,” *arXiv preprint arXiv:2506.23283*, 2025.
- [95] K. Li, X. Li, Y. Wang, Y. He, Y. Wang, L. Wang, and Y. Qiao, “Videomamba: State space model for efficient video understanding,” *arXiv preprint arXiv:2403.06977*, 2024.
- [96] M. H. Erol, A. Senocak, J. Feng, and J. S. Chung, “Audio mamba: Bidirectional state space model for audio representation learning,” *arXiv preprint arXiv:2406.03344*, 2024.
- [97] J. Ma, F. Li, and B. Wang, “U-mamba: Enhancing long-range dependency for biomedical image segmentation,” *arXiv preprint arXiv:2401.04722*, 2024.
- [98] J. Liu, M. Liu, Z. Wang, P. An, X. Li, K. Zhou, S. Yang, R. Zhang, Y. Guo, and S. Zhang, “Robomamba: Efficient vision-language-action model for robotic reasoning and manipulation,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 40 085–40 110, 2024.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: Yes, the abstract and introduction accurately reflect the paper's contributions and scope.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Yes, our paper discussed the limitations of the work in the Appendix A.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper is application-oriented and does not contain any theoretical results, assumptions, or formal proofs. The focus is on model design and benchmark evaluation rather than theoretical development.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides comprehensive details necessary for reproducing the main experimental results. In Section 3, we describe the model architecture and in Section 4.3, we elaborate on the training setup.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We will release the code and dataset on publication.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper specifies experimental details in Section 3 (Method) and Section 4.3 (Implementation). Furthermore, Section 4.5 (Ablation Study) details the configurations and modifications used in component-level experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Statistical significance analysis is generally not reported in referring image segmentation literature. It is widely considered unnecessary and is not part of the standard evaluation protocol in this field.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Yes, we provide sufficient information in Section 4.3 (Implementation).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We have reviewed the NeurIPS Code of Ethics and confirm that our research fully conforms to its principles.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Yes, we discuss broader impacts in the Appendix B.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our paper does not involve the release of models or datasets with high risk of misuse. The released components are task-specific and do not pose safety or dual-use concerns.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All external assets used in our work—including pre-trained models (e.g., Swin Transformer-B and BERT-B) and datasets—are properly cited in the paper. We explicitly follow their respective licenses: Swin Transformer (MIT License) and BERT (Apache 2.0 license). For all datasets (RefCOCO family), we ensured compliance with their original license terms (Apache 2.0 license).

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We will publicly release assets (code, checkpoints, and dataset) on publication. Full documentation will be provided alongside the assets to ensure reproducibility and ease of use.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing or human subject research in its core contribution.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The work involves no new human subjects or data collection requiring IRB review.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: A large language model (Qwen) was used solely during the dataset construction phase to assist in generating textual annotations for aRefCOCO. The LLM does not play any role in the core methodology, model design, or training.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Limitations

While our method achieves strong performance across various benchmarks, several limitations must be acknowledged. First, the model’s ability to process extremely complex or ambiguous expressions remains limited. Although we evaluated it primarily on datasets like RefCOCO family and aRefCOCO, these datasets may not fully represent the diversity of real-world scenarios. Additionally, for dataset construction, we rely on a large language model to generate textual descriptions. While this approach ensures a broad range of expressions, it may inadvertently introduce bias, even after meticulous filtering.

B Broader Impacts

The work presented here has the potential for both positive and negative societal impacts. On the positive side, our method’s ability to understand complex referring expressions and improve image segmentation could significantly benefit fields such as medical imaging, robotics, and accessibility technologies—particularly for individuals with disabilities. On the negative side, the model may inherit biases from pre-trained encoders (e.g., Swin or BERT), potentially leading to stereotypical associations in referring expressions (e.g., linking gender or social roles to certain objects), thereby amplifying societal biases in segmentation outcomes.

C More Related Work

From Traditional Segmentation to Language-Based Segmentation. Traditional segmentation has long focused on delineating object regions from visual cues alone, employing various methods such as CNN [67–72], Vision Transformers [73–78], and Diffusion Models [79–86]. Moving beyond purely visual cues, **language-based segmentation** introduces linguistic grounding, enabling segmentation guided by natural language. Within this field, *referring image segmentation* (RIS) aims to localize and segment objects described by textual expressions, yet many methods still treat it as a shallow keyword-matching task, leaving ambiguous expressions unresolved. *Open-vocabulary segmentation* (OVS), in contrast, emphasizes category generalization and broad semantic coverage through large-scale vision–language pretraining [61–63, 87–90]. More recently, *reasoning segmentation*, introduced by LISA [19] extends beyond RIS by requiring complex reasoning or world knowledge, leveraging large language models to perform such inference. Our work, however, remains within the RIS scope: rather than relying on external knowledge, we focus on resolving the referential ambiguities that an ideal RIS model *should* handle but current approaches *fail* to address.

Mamba and Its Broader Applications. Recently, **Mamba** has been actively explored across unimodal and multimodal learning [91–93], owing to its ability to model long-range dependencies with linear-time efficiency. It has since been adopted as a general-purpose backbone in diverse downstream tasks. In the visual domain, Mamba has shown promise in video understanding and spatiotemporal modeling [94, 95], where its state-space formulation enables efficient temporal integration while preserving spatial fidelity. Beyond vision, it has been applied to audio and speech modeling [96], biomedical image analysis [97], and robotics with vision–language–action models [98], demonstrating strong adaptability across modalities. These developments highlight Mamba’s growing influence as a unified sequence modeling paradigm. In this work, we further leverage its long-context sequential modeling capacity and scan-then-update dynamics—mirroring the saccade–fixation process of human perception—to resolve referential ambiguities in referring image segmentation.

D More Details of aRefCOCO Dataset

D.1 Generation, Filtering and Validation

D.1.1 Generation Prompts

In order to ensure the richness and diversity of object descriptions while adhering to the fundamental requirement of single-object localization in RIS tasks, we used a prompt template for Qwen-VL-Max as shown in Appendix D.1.1.

Generation Prompts

Role

You are a professional assistant, proficient in object recognition and description. Your task is to provide a clear and focused description of the main object inside the red box in the image. You are given the category information of the main object inside the red box as '{target}'. The description should be based on the provided target information, and you should generate a more detailed description with your visual analysis of the image. Please note that there is no visible box in the actual image.

Target Category Information

target category information: '{target}'

Skills

1. [Important!] Please only focus on the main object, i.e., '{target}', and ignore the background or irrelevant details.
2. [Important!] Do not assume or include any information that is not present in the image content.
3. [Important!] Provide 5 independent English sentences, each of which can independently identify '{target}' without relying on other sentences. Each sentence should describe a unique identifying feature.
4. [Important!] Use the subject '{target}' from the target category information, and ensure you are describing a single object '{target}'! Others should be able to locate only one object based on your sentences.
5. [Important!] Use singular language to ensure the description focuses only on one main object. Do not use plural words like 'they' as the subject, but you are allowed to use pronouns as the subject.
6. [Important!] If you need to describe based on objects around '{target}', please keep the subject as '{target}'.

D.1.2 Filtering

We first pre-screen all descriptions with GPT-4o for the two patterns: "object-distracting" (i.e., containing multiple nouns) and "category-implicit" (i.e., where the subject is a pronoun). Items that match either pattern are retained. Then, to verify the remaining pairs meet the basic rules we set for annotation, GPT-4o is involved as an evaluator. The prompt is as shown in Appendix D.1.2.

Filtering Prompts

Role:

You are a professional visual-language evaluator. Your task is to verify whether each given description sentence is a faithful, grounded, and rule-abiding description of the target object **inside the red box** in the provided image.

Information:

The target object belongs to the category: {target}.

The description was generated under the following constraints:

1. It must describe only the main object ('{target}'), ignoring the background or other irrelevant elements.
2. It must not include objects that are not in the whole image, but objects out of box are allowed.
3. Each sentence must independently identify the target.
4. It can contain objects beyond the target.
5. It must use singular phrasing, and refer to the object using either its category label ('{target}') or singular pronouns.

You will be given 5 sentences. Your task is to verify, for each sentence, whether it satisfies **all** the above requirements based solely on the image content.

Sentences to Verify:

{Sentences}

Instructions:

1. For each sentence, respond with ****"yes"**** or ****"no"****. if "no", follow it by a ****brief reason****.
2. If any rule is violated or the content is ungrounded, answer "no" and specify which rule(s) are broken.
3. Do not use external knowledge or assumptions. Base your judgment strictly on what is visible in the image.
4. Provide one answer per sentence, in order. Each answer must be on its own line.
5. Do not add commentary outside the answers. Return only the answers.
6. Ensure the number of answers equals 5.

D.1.3 Validation

Manual review was conducted by at least two annotators with RIS domain expertise. Each pair was independently validated to confirm (1) whether the description uniquely refers to a single object in the image, and (2) whether the expression obeys our referential ambiguity criteria while remaining image-groundable. In borderline cases, annotators reached consensus through discussion, and a third reviewer was consulted when needed.

D.2 Comparison with Original Descriptions

We conduct quantitative comparisons to highlight the distinct features of aRefCOCO compared to the original descriptions, as visualized in Fig. 6. Specifically, the comparisons focus on the following aspects:

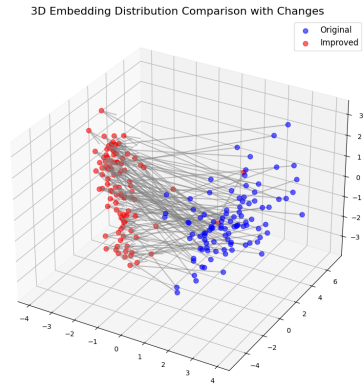
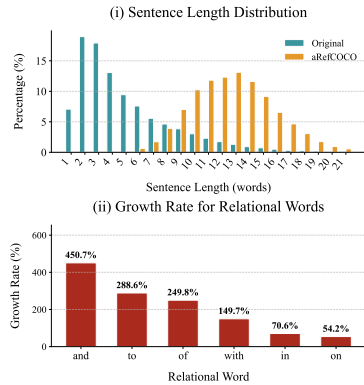
Description Length. As shown in Fig. 6a(i), compared to the original descriptions in RefCOCO and RefCOCOg, aRefCOCO contains a higher proportion of longer sentences and significantly reduces the frequency of short, phrase-like descriptions. This shift towards more complex sentence structures underscores the value of aRefCOCO as an extension to the RefCOCO series, making it particularly relevant for real-world RIS tasks that involve longer and more intricate sentence constructions.

Relational Words. We calculated the growth rates of the most frequent words in the original and aRefCOCO descriptions, focusing on those strongly correlated with the relationships between targets and distractors. The results in Fig. 6a(ii) show a significant increase in these words, particularly prepositions like “and”, “to”, and “of”, which grew by 450.7%, 288.6%, and 249.8%, respectively. These growths contribute to more structurally and semantically complex sentences, enhancing spatial and relational expressions and enriching the semantic content in aRefCOCO descriptions.

Text Embedding Distribution. To further examine the semantic differences between the original and aRefCOCO descriptions, we visualize their text embeddings in a reduced 3D space using PCA, as shown in Fig. 6b. The figure illustrates a clear shift in embedding distributions, where the aRefCOCO (red) embeddings are consistently displaced from their original (blue) counterparts. This systematic divergence indicates that the rewritten descriptions introduce substantial semantic variations rather than minor lexical edits.

E More Visualization Results

In this section, we provide more visualization results of our **SaFiRe**, ReMamber [9] and PSALM [52] on aRefCOCO for further comparison, as shown in Fig. 7.



(a) Sentence-level statistics comparison

(b) 3D text embedding distribution comparison

Figure 6: Comparison of Sentence-Level Statistics and Text Embedding Distributions Between the Original Descriptions and aRefCOCO.

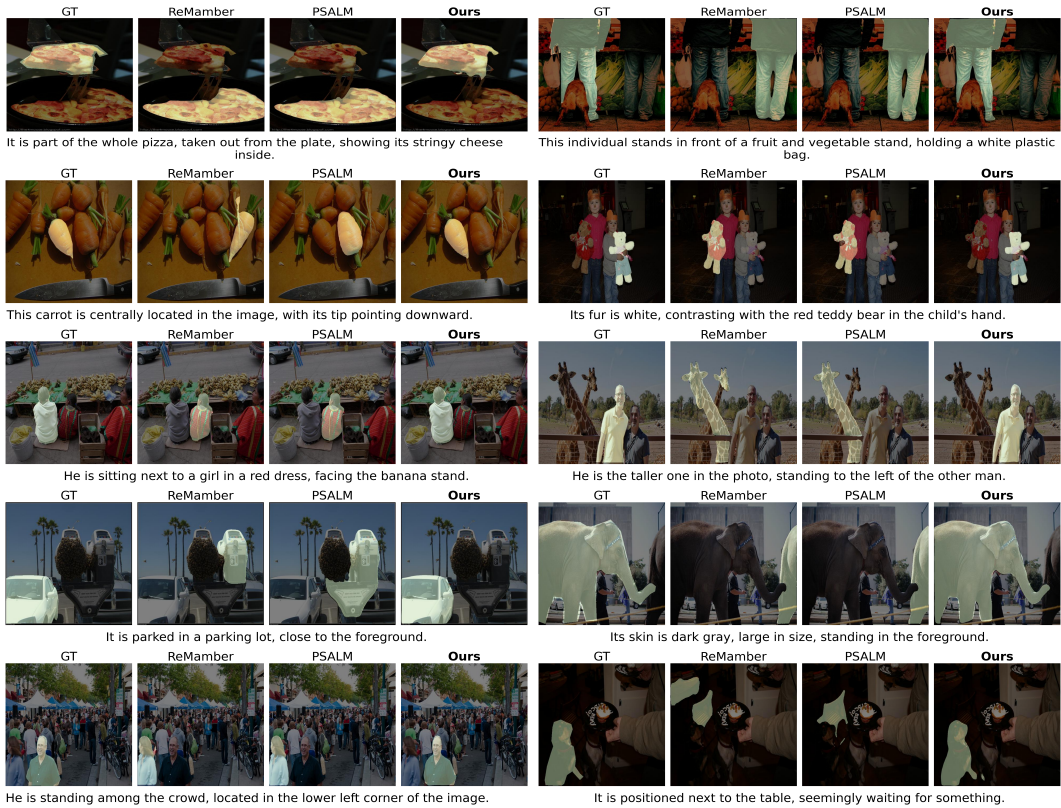


Figure 7: More Results on aRefCOCO (Part 1).

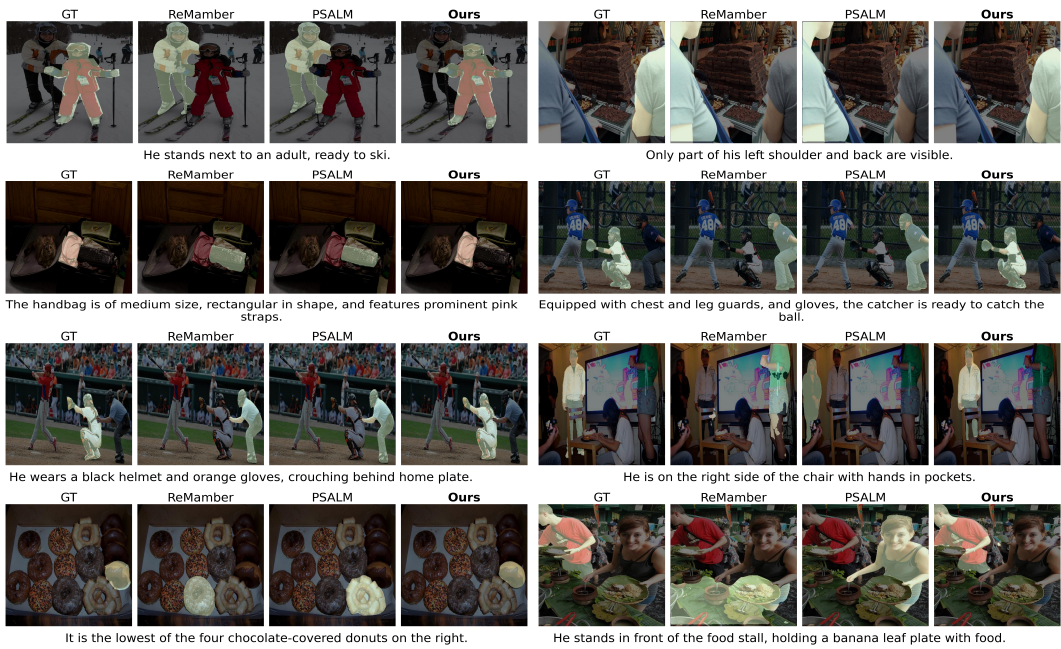


Figure 7: More Results on aRefCOCO (Part 2, continued).