Visual Expression for Referring Expression Segmentation

Anonymous ACL submission

Abstract

Referring expression segmentation aims to segment a target object precisely in the image by referencing to a given linguistic expression. Since the network predicts based on the reference information that guides the network on which regions to pay attention, the capacity of this guidance information has a significant impact on the segmentation result. However, most existing methods rely on linguistic context-based tokens as the guidance elements, which are limited in providing the visual understanding of the fine-grained target regions. To address this issue, we propose a novel **M**ulti-**E**xpression Guidance framework for **R**eferring **E**xpression **S**egmentation, **MERES**, which enables the network to refer to the visual expression tokens as well as the linguistic expression tokens to complement the linguistic guidance capacity by effectively providing the visual contexts of the fine-grained target regions. To produce the semantic visual expression tokens, we introduce a visual expression extractor that adaptively selects the useful visual information relevant to the target regions from the image context and allows the visual expression to capture the richer visual contexts. The proposed module strengthens the adaptability to the diverse image and language inputs, and improves visual understanding of the target regions. Our method consistently shows strong performance on three public benchmarks, where it surpasses the existing state-of-the-art methods.

1 Introduction

Referring expression segmentation (RES) [22, 40, 45, 44, 26] is one of the challenging vision-language tasks [6, 69, 20, 37], and can be applied in various applications such as human-robot interaction and the object retrieval. Given an image and a natural language expression describing a target object within the image, one of the key points in this task is for the network to precisely segment the target object regions from the image by referring to the given expression. Unlike the single modal segmentation [60, 31, 27] based on fixed categories, the RES treats the free-form language expressions. For instance, the language expression can be given as a word that represents a single attribute, such as *"left"*, or as a phrase or sentence that represents more than one attribute, such as *"pink shirts on the sofa"*. On the other hand, the image context contains more diverse information of the target object beyond the location, color and relationships, such as the fine-grained region information with irregular shape that is difficult to describe in the language expression. In this paper, we address the limitation of the linguistic expression, which contains only some part of the target region information.

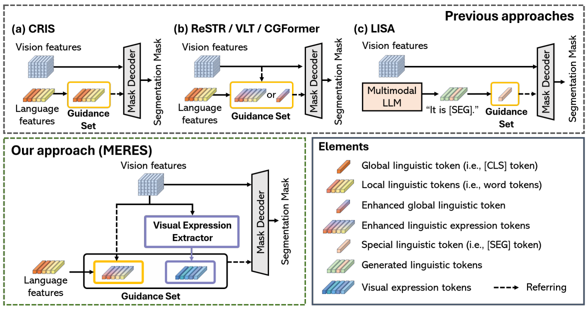

Figure 1: Guidance set comparison of our approach and previous approaches [68, 32, 15, 62, 38]. Unlike these approaches, our approach allows visual expression tokens as well as linguistic expression tokens to be used as guidance elements, to complement linguistic guidance capacity. Extended figure is available in Appendix A.

Existing methods [68, 32, 61] have focused on the multi-modal fusion, which enables vision features to effectively refer to the language features. Some studies [24, 15, 62] have focused on improving the comprehension for the linguistic expression by allowing language features to refer to the vision features via the language-vision cross-attention mechanism. These methods successfully address the ambiguity of the language expression by obtaining the enhanced linguistic features. More recent studies [38, 57] employ large language mod-

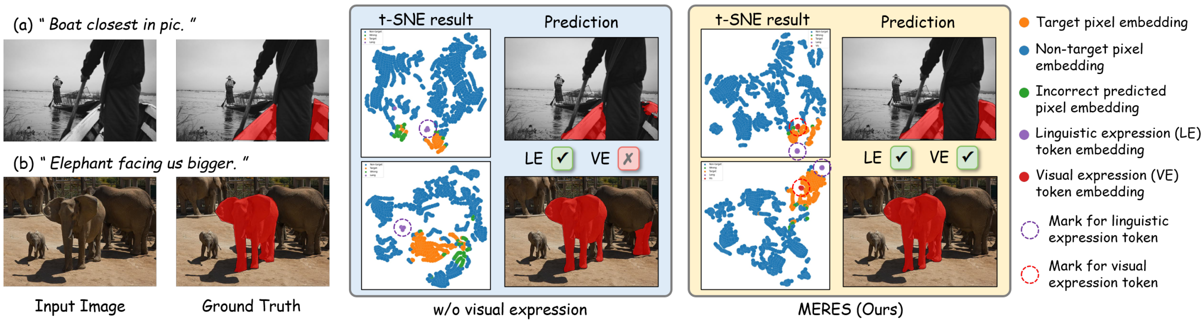

Figure 2: t-SNE and qualitative results of the ablation method and the proposed method. In t-SNE results, our VE tokens help to better cluster the target pixel embeddings, whereas LE tokens of the ablation method cannot sufficiently cluster the target pixels. In segmentation results, due to the lack of visual understanding of the fine-grained target regions, the ablation method guides the network to segment only some part of the target regions (*i.e.*, boat) or segment even non-target regions (*i.e.*, other elephant's leg). In contrast, our method shows robust segmentation by complementing the linguistic guidance capacity and providing visual contexts of the target regions.

els (LLMs) [65, 11, 63, 76] to improve the understanding of the language expression via LLM's immense knowledge, and exploit the generated language tokens in the segmentation network. However, as displayed in Figure 1, all these methods rely on the linguistic context-based tokens to guide the network to segment the target regions. Since these tokens are insufficient to capture the visual contexts, these linguistic-based tokens are limited in providing the visual understanding that helps guide the network to the target areas composed of the fine-grained regions with different visual characteristics. For example, in Figure 2, the network guided by only linguistic-based tokens segments only part of the target regions (*i.e.*, (a)) or segments even non-target regions (*i.e.*, (b)).

To tackle this issue, we focus on producing the visual expression tokens that can complement the linguistic information by effectively providing the visual contexts of the target regions; the set of such tokens that provide the target region information to the network is called *Guidance Set* in this paper. The role of the guidance set is to guide the network on which regions to focus its attention, because the network predicts target regions based on the guidance information. Thus, we explore the capability of the guidance set, which has a significant impact on the segmentation results in the referring expression segmentation task.

In this paper, we propose a novel Multi-Expression guidance framework for Referring Expression Segmentation, MERES, which enables the network to refer to the advanced guidance set composed of the visual expression tokens as well as the linguistic expression tokens. The proposed framework is distinct from previous studies in that we produce the visual expression tokens to enhance the capacity of the guidance set and avoid relying only on the linguistic guidance, as illustrated in Figure 1. Our visual expression tokens complement the linguistic guidance capacity by effectively providing the visual contexts of the target regions.

To produce the semantic visual expression, we design a visual expression extractor from the terms of two points: (1) It needs to selectively exploit the semantic information relevant to the target regions from the image context that contains both target and non-target region information. (2) It needs to consider the rich visual contextual information of the target regions. For (1), the relevance to the given linguistic expression can be used as a cue to identify some degree of target regions in the image context, but if there is insufficient target information in the linguistic cue, the useful visual information may not be selected due to the weak relevance to the linguistic features. To prevent this over-reliance on high relevance to the linguistic cue at the selection step, our module *adaptively* selects the semantic information related to the target regions from the image context. This strengthens the adaptability to diverse language and image inputs for robust segmentation. For (2), our module allows the visual expression tokens to consider richer visual contexts by leveraging the global-local linguistic cues (*i.e.*, sentence-level and word-level cues), where each of linguistic cues has different contextual information, and by acquiring the relationship between each visual token. This improves the visual understanding of the fine-grained target regions.

We demonstrate the effectiveness of the proposed approach on three public RES benchmarks. In particular, our approach outperforms previous state-of-the-art methods on all of three datasets. Our contributions are summarized as follows:

- We propose a novel Multi-Expression guidance framework for Referring Expression Segmentation, MERES, which enables the network to refer to the advanced guidance set composed of visual expression tokens as well as linguistic expression tokens, to complement linguistic guidance capacity.

- To produce more semantic visual expression tokens, we introduce a visual expression extractor that adaptively selects the useful information related to the target regions from image context and allows the visual expression to consider the richer visual contexts. Our module enhances the adaptability to diverse image and language inputs, and improves visual understanding of the target regions.

- Our method consistently shows strong performance and surpasses previous state-of-the-art methods on three widely-used RES datasets.

2 Related Works

Referring Expression Segmentation. Different from the unimodal segmentation [56, 70, 79] based on predefined categories, referring segmentation addresses the unrestricted language expression. Recent researches [32, 68, 71, 61] have explored on the better multi-modal fusion for this task. Other recent studies [59, 3, 15] have incorporated the visual information into the language features. KWAN [59] captured the visual context features for keywords and concatenated them with the vision features. STEP [3] extracted the visual-attended text representation to obtain the heatmap. VLT [15] improved the comprehension for the language expression and captured the enhanced language features by referring to the vision features. ReLA [43] and DMMI [23] proposed the generalized RES datasets that contain multi-target and no-target samples. JMCELN [26] used learnable embeddings to adaptively obtain multi-modal contextual information. CGFormer [62] exploited the linguistic tokens for grouping visual features. SADLR [72] iteratively updated the segmentation mask and the global linguistic features.

Unlike these approaches, as shown in Figure 9 of Appendix, our approach focuses on producing the visual expression tokens to complement the linguistic guidance capacity, which can effectively provide the visual understanding of the target regions.

Token Selection. Recent studies have exploited the top-k method for the token selection to flexibly select the useful tokens in various tasks. TS-ViT [78] proposed a drop-in token selection method to improve the selectivity of the self-attention and enhance the robustness of the transformer models. For patch selection in large images, DPS [12] exploited the top-k method to aggregate information from the different patches in a flexible manner. For video object segmentation, HMMN [58] proposed a top-k guided memory matching method, resulting in efficient and robust fine-scale memory matching. MiVOS [9] proposed a top-k filtering scheme for the attention-based memory read operation. TS2-Net [47] proposed a token shift and selection transformer that dynamically selects informative tokens in both temporal and spatial dimensions on text-to-video retrieval. PPMN [19] proposed a pixel-noun matching network using top-k selection to endow noun features with stronger discriminative ability on panoptic segmentation.

We thus leverage the top-k selection method in our visual expression extractor to prevent over-reliance on high relevance to the linguistic cues at the selection step by adaptively selecting the useful visual information associated with target regions on referring expression segmentation.

3 Proposed Method

We propose a novel multi-expression guidance framework on referring expression segmentation, MERES, to avoid relying on linguistic guidance. Figure 3 shows the overall framework. We first describe the vision and language feature extraction (Sec.3.1), and then introduce a visual expression extractor (Sec.3.2). Finally, we explain a segmentation decoder (Sec.3.3).

3.1 Vision and Language Feature Extraction

Given the input image $\mathcal{I}$ and the linguistic expression $\mathcal{Q}$ that consists of $T-1$ words, the vision encoder extracts the vision features $\mathbf{F}_i \in \mathbb{R}^{H_i W_i \times C_i}$ at each stage $i \in \{1, 2, 3, 4\}$ and the language encoder extracts the linguistic expression tokens $\mathbf{Q}_l = [\mathbf{w}_{cls}, \mathbf{w}_1, ..., \mathbf{w}_{T-1}] \in \mathbb{R}^{T \times D}$. Note that H_i, W_i, C_i and D denote the height, width, channel dimension of the feature maps at the i^{th} vision stage, and the channel dimension of the linguistic features. The first token $\mathbf{w}_{cls}$ of linguistic expression features is a special `[CLS]` token, which is the global representation that understands the linguistic expression at the sentence level.

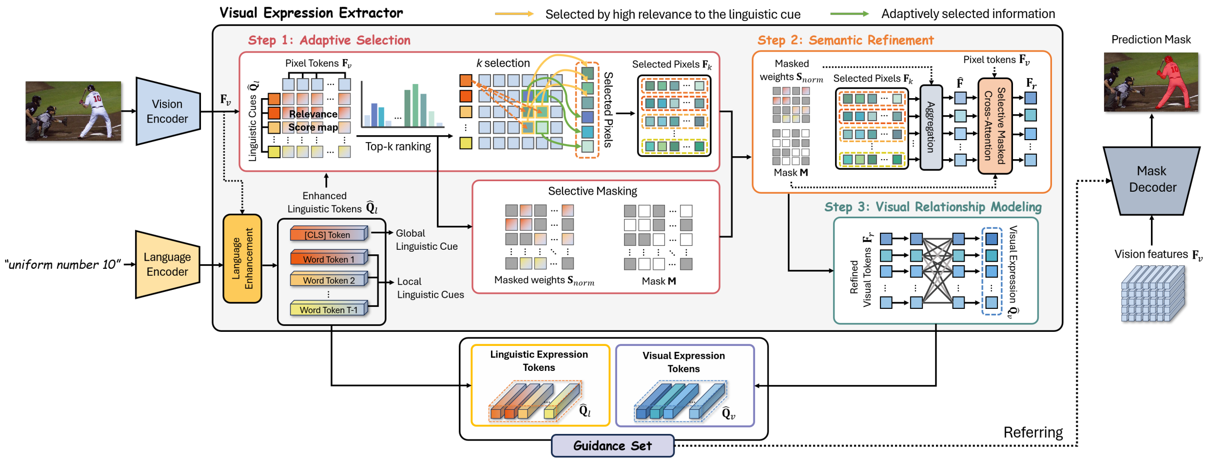

Figure 3: Overview of the proposed framework. Our method improves the robustness of the guidance set capacity by producing the visual expression. The visual expression extractor produces the visual expression tokens via three steps: the adaptive selection, the semantic refinement and the visual relationship modeling.

3.2 Visual Expression Extractor

To improve the guidance capability, we produce the visual expression that contains the visual semantic contexts related to the target object. As illustrated in Figure 3, the visual expression extractor consists of three steps: i) adaptive selection, ii) semantic refinement, iii) visual relationship modeling.

Adaptive selection. This step leverages the global-local linguistic cues to consider both the comprehensive and specific attribute contexts for the rich contextual information, as each linguistic cue captures the different contextual embedding. In this step, the linguistic expression tokens are first enhanced by the cross-attention layers using the vision features as key-value pairs to improve the comprehension for the language contexts. Then, the vision features $\mathbf{F}_v(=\mathbf{F}_4) \in \mathbb{R}^{N\times C}$ and the enhanced global-local linguistic tokens $\widehat{\mathbf{Q}}_l$ are embedded into the joint embedding space by the linear projection ϕ, where N is the total number of pixels. This process is formulated as follows:

$$\mathbf{X} = \phi^{\mathcal{V}}(\mathbf{F}_v)\ ,\ \mathbf{Y} = \phi^{\mathcal{L}}(\widehat{\mathbf{Q}}_l)\ , \tag{1}$$

After that, the relevance score map $\mathbf{S} \in \mathbb{R}^{T\times N}$ between the vision pixel tokens and the linguistic tokens is computed to rank, and the pixel tokens are selected by the top-k method as follows:

$$\mathbf{S} = \mathcal{S}(\mathbf{X}, \mathbf{Y})\ ,\ \mathbf{S}_k = \mathcal{K}(\mathbf{S})\ , \tag{2}$$

where $\mathcal{S}$, K, $\mathbf{S}_k \in \mathbb{R}^{T\times K}$ and $\mathcal{K}$ denote the cosine similarity function, the number of the selected pixels, the set of the selected pixel index lists per linguistic token, and the top-k operation. As shown in Figure 3, the top-k ranked pixel tokens $\mathbf{F}_k \in \mathbb{R}^{K\times D}$ are used to produce the visual expression tokens. Even if there is insufficient target region information in the linguistic cues, the top-k method enables to adaptively select the semantic visual information that has weak relevance to linguistic cues but is useful for robust segmentation. This adaptive selection prevents over-reliance on the high relevance to the linguistic cues.

To prevent the high relevance scores between the linguistic cues and the incorrect regions, the relevance score map $\mathbf{s} \in \mathbb{R}^{1\times N}$ of the global linguistic token is supervised by the pixel contrastive loss:

$$\mathcal{L}_c = \begin{cases} -\log(\sigma(\mathbf{s}_j/\tau)) & if\ j \in \mathcal{Z}^+ \\ -\log(1-\sigma(\mathbf{s}_j/\tau)) & if\ j \in \mathcal{Z}^- \end{cases}, \tag{3}$$

where $\mathcal{Z}^+$ and $\mathcal{Z}^-$ denote the set of the relevant pixels and irrelevant pixels for the target regions. τ is a learnable temperature, and σ is a sigmoid function. The pixel contrastive loss encourages that the relevant pixels are embedded closer together for high relevance score and the irrelevant pixels are embedded far apart for low relevance score.

Semantic refinement. The selected useful tokens are passed to the semantic refinement step. Rather than simply aggregating the selected information, it is more effective to refine the selected information as the network adaptively captures the useful information from the selected information to produce more semantic visual expression tokens. In the semantic refinement step, the aggregated visual tokens $\mathbf{F}_a \in \mathbb{R}^{T\times D}$ are first obtained by the top-k weighted average pooling, as follows:

$$n \in \{1, 2, ..., N\},\ t \in \{1, 2, ..., T\}, \tag{4}$$

$$\mathbf{M}_n^t = \begin{cases} 0 & n \in \mathbf{S}_k^t \\ -\infty & n \notin \mathbf{S}_k^t \end{cases}, \mathbf{M} = [\mathbf{M}^1, ..., \mathbf{M}^T], \tag{5}$$

$$\mathbf{S}_{norm} = \texttt{Reshape}(\texttt{softmax}(\mathbf{S}+\mathbf{M})), \tag{6}$$

$$\mathbf{F}_a = \frac{1}{K}\sum^{K}(\mathbf{S}_{norm} \odot \texttt{Repeat}(\mathbf{F}_v, T)), \tag{7}$$

where $\odot$ and $\mathbf{M} \in \mathbb{R}^{T \times N}$ denote the element-wise multiplication and the top-k selective mask that masks the non top-k ranked scores. $\texttt{Repeat}(f, x)$ indicates repeating the feature f x times to expand the shape. Since the top-k selection is discrete, the normalized top-k score map $\mathbf{S}_{norm} \in \mathbb{R}^{T \times N \times 1}$ is obtained by normalizing the whole relevance score map $\mathbf{S}$ combined with the top-k selective mask $\mathbf{M}$.

Then, the refined visual tokens $\mathbf{F}_r \in \mathbb{R}^{T \times D}$ are obtained by refining each aggregated visual token via the selective masked cross-attention mechanism to dynamically capture the useful semantic information from the top-k selected pixels, as follows:

$$\widehat{\mathbf{F}} = \texttt{MCA}(\mathbf{F}_a, \mathbf{F}_v, \mathbf{M}) + \mathbf{F}_a,\ \mathbf{F}_r = \texttt{MLP}(\widehat{\mathbf{F}}) + \widehat{\mathbf{F}}, \quad (8)$$

where $\texttt{MCA}$ denotes the masked cross-attention, and $\widehat{\mathbf{F}}$ is the intermediate features.

Visual relationship modeling. The visual expression tokens $\widehat{\mathbf{Q}}_v = [\mathbf{v}_{cls}, \mathbf{v}_1, ..., \mathbf{v}_{T-1}] \in \mathbb{R}^{T \times D}$ are produced by considering the visual relationship to mutually complement each visual token's information and capture the visual contextual information, improving the visual understanding of the fine-grained target regions, formulated as:

$$\widehat{\mathbf{Q}} = \texttt{MHSA}(\mathbf{F}_r) + \mathbf{F}_r\ ,\ \widehat{\mathbf{Q}}_v = \texttt{MLP}(\widehat{\mathbf{Q}}) + \widehat{\mathbf{Q}}\ , \quad (9)$$

where $\texttt{MHSA}$ and $\widehat{\mathbf{Q}}$ indicate the multi-head self-attention, and the intermediate features.

3.3 Segmentation Decoder

To segment the target region, the decoder leverages the guidance set $\mathcal{G} = \{\widehat{\mathbf{Q}}_l, \widehat{\mathbf{Q}}_v\}$ composed of the enhanced linguistic expression tokens and the visual expression tokens. The decoder can focus its attention on more precise target regions due to the enhanced guidance for the visual understanding of target regions. At each decoder stage, the cross-attention layer, which uses the vision features as the query and the guidance tokens as the key-value, is employed to highlight the target regions. The vision decoder features are then upsampled and concatenated with the corresponding vision encoder features to feed into the next decoder stage. The final segmentation map is projected to a binary class mask by a linear projection layer. The binary cross-entropy loss is used for the network training.

4 Experiments

4.1 Implementation Details

Experimental settings. The vision encoder is Swin-B [48] initialized with the pre-trained weight on ImageNet-22K [35], and the language encoder is BERT-base [14] initialized with the official pre-trained weight of the uncased version. The decoder was randomly initialized. We trained models for 40 epochs with 16 batch size on 24G RTX3090 GPUs.

Datasets. RefCOCO [73] and RefCOCO+ [73] are widely utilized datasets for referring image segmentation. RefCOCO contains 19,994 images with 142,209 language expressions for 50,000 objects, and RefCOCO+ contains 19,992 images with 141,564 expressions for 49,856 objects. The expressions in RefCOCO+ do not include words about absolute locations, which makes it more challenging than RefCOCO. G-Ref [51, 52] is also a commonly used dataset, which contains 26,711 images with 104,560 language expressions for 54,822 objects. G-Ref, which is the most challenging dataset, has more complex and longer expressions than RefCOCO and RefCOCO+.

Evaluation metrics. Following previous works, we adopted the overall intersection-over-union (oIoU), mean intersection-over-union (mIoU), and precision at 0.5, 0.7 and 0.9 thresholds. More details for settings and metrics are in Appendix C.

4.2 Comparison with State-of-the-Arts

In Table 1, we evaluated our approach with previous state-of-the-art methods on three public benchmarks for referring expression segmentation. Our method consistently showed strong performance on all evaluation splits of all datasets, and outperformed other existing methods on three benchmarks. Compared to VLT [15], which leverages the enhanced linguistic features as the guidance set elements, our MERES improved oIoU performance by 2.39%, 2.01% and 2.34% on each split of RefCOCO, respectively. Compared to the recent state-of-the-art method, CGFormer [62], our model showed 2.16%, 1.08% and 2.71% higher oIoU performance on each split of RefCOCO+. Compared to the other recent method, VG-LAW [61], our model achieved significant mIoU improvements of 2.49% and 2.84% on each split of G-Ref, the most challenging dataset. These results demonstrate the effectiveness of our approach.

In addition, we compared on the generalization setting to further validate the generalization ability in Table 2. Our MERES surpassed the existing methods and consistently showed performance improvements on both seen and unseen sets. These results suggest that our method has a better generalization ability than other methods in this task.

	Method	Venue	Vision Encoder	Language Encoder	RefCOCO val	RefCOCO test A	RefCOCO test B	RefCOCO+ val	RefCOCO+ test A	RefCOCO+ test B	G-Ref val$_{(U)}$	G-Ref test$_{(U)}$	G-Ref val$_{(G)}$
mIoU	MCN [50]	CVPR '20	DarkNet53	Bi-GRU	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40	-
	LTS [28]	CVPR '21	DarkNet53	Bi-GRU	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25	-
	CRIS [68]	CVPR '22	CLIP-R101	CLIP	70.47	73.18	66.10	62.27	68.08	53.60	59.87	60.36	-
	JMCELN [26]	EMNLP '23	CLIP-R101	CLIP	74.40	77.69	70.43	66.99	72.69	57.34	64.08	64.99	-
	PVD [10]	AAAI '24	Swin-B	BERT-base	75.07	77.29	70.13	64.39	69.15	57.19	63.22	63.89	61.74
	VG-LAW [61]	CVPR '23	ViT-B	BERT-base	75.05	77.36	71.69	66.61	70.30	58.14	65.36	65.13	-
	MERES (Ours)	-	Swin-B	BERT-base	**76.97**	**78.89**	**73.63**	**68.63**	**73.88**	**61.94**	**67.85**	**67.97**	**65.86**
oIoU	CMPC [25]	CVPR '20	ResNet101	LSTM	61.36	64.53	59.64	49.56	53.44	43.23	-	-	49.05
	ReSTR [32]	CVPR '22	ViT-B	Transformer	67.22	69.30	64.45	55.78	60.44	48.27	54.48	-	-
	LAVT [71]	CVPR '22	Swin-B	BERT-base	72.73	75.82	68.79	62.14	68.38	55.10	61.24	62.09	-
	CoupAlign [77]	NeurIPS '22	Swin-B	BERT-base	74.70	77.76	70.58	62.92	68.34	56.69	62.84	62.22	-
	VLT [15]	TPAMI '23	Swin-B	BERT-base	72.96	75.96	69.60	63.53	68.43	56.92	63.49	66.22	62.80
	ReLA [43]	CVPR '23	Swin-B	BERT-base	73.82	76.48	70.18	66.04	71.02	57.65	65.00	65.97	62.70
	DMMI [23]	ICCV '23	Swin-B	BERT-base	74.13	77.13	70.16	63.98	69.73	57.03	63.46	64.19	61.98
	SADLR [72]	AAAI '23	Swin-B	BERT-base	74.24	76.25	70.06	64.28	69.09	55.19	63.60	63.56	61.16
	CGFormer [62]	CVPR '23	Swin-B	BERT-base	74.75	77.30	70.64	64.54	71.00	57.14	64.68	65.09	62.51
	MERES (Ours)	-	Swin-B	BERT-base	**75.35**	**77.97**	**71.94**	**66.70**	**72.08**	**59.85**	**65.78**	**66.93**	**63.49**

Table 1: Performance comparison with the existing state-of-the-art methods on three widely-used referring expression segmentation benchmarks. (U): UMD split. (G): Google split. Best score is in **bold**.

Method	Vision Encoder	Language Encoder	val$_{(U)}$ *seen*	val$_{(U)}$ *unseen*	test$_{(U)}$ *seen*	test$_{(U)}$ *unseen*	val$_{(G)}$ *seen*	val$_{(G)}$ *unseen*
CRIS [68]	CLIP-R101	CLIP	58.64	42.63	59.68	38.88	42.36	32.84
LAVT [71]	Swin-B	CLIP	60.16	42.33	60.37	41.38	57.33	40.43
CGFormer [62]	Swin-B	CLIP	65.60	46.11	65.67	42.31	62.85	45.05
MERES (Ours)	Swin-B	CLIP	**66.52**	**46.74**	**66.93**	**43.06**	**63.61**	**46.01**

Table 2: Comparison for generalization setting [62] on G-Ref using mIoU. Details for setting is in Appendix E.

Guidance Elements	RefCOCO val P@0.5	P@0.7	P@0.9	mIoU	oIoU	G-Ref val$_{(U)}$ P@0.5	P@0.7	P@0.9	mIoU	oIoU
LE	84.73	75.49	34.87	74.61	72.85	72.77	59.90	22.86	62.52	61.59
Enhanced LE	85.46	76.22	36.04	75.10	73.56	74.02	61.28	24.55	64.35	63.68
VE	86.38	77.82	36.90	75.84	74.52	74.89	63.03	26.33	66.31	65.45
Enhanced LE + All pixels	86.17	77.40	36.73	75.65	74.36	74.85	62.77	25.91	66.02	65.27
Enhanced LE + VE	**86.71**	**78.30**	**37.24**	**76.97**	**75.35**	**76.13**	**64.60**	**27.87**	**67.85**	**66.93**

Table 3: Main ablation for the effectiveness of our approach. LE: Linguistic Expression tokens. VE: Visual Expression tokens. Our full model is in gray.

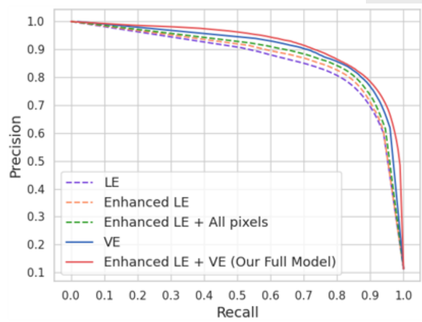

Figure 4: Precision-Recall curves of our model and the ablation models on RefCOCO+ testA.

4.3 Ablation Studies

4.3.1 Effectiveness of Proposed Framework

In Table 3, we conducted experiments to validate the effectiveness of using the visual expression tokens as well as the linguistic expression tokens as the elements of the guidance set. All ablation models are based on our network. Compared to 'LE only' method that uses only the pure language encoder features $\mathbf{Q}_l$ as guidance elements, 'Enhanced LE' method, which uses only the enhanced linguistic tokens $\widehat{\mathbf{Q}}_l$ as guidance elements, showed better performance on each dataset. This suggests that the enhancement of the language features by referring to the visual information helps to improve the comprehension for the meaning of the language expression context. Compared to these two methods, our full method showed remarkable improvements on both datasets. These results indicate that linguistic guidance capacity is insufficient to provide the visual understanding of the fine-grained target regions, and using visual expression tokens as guidance elements can effectively complement the linguistic guidance capacity.

Furthermore, 'VE only' method (row3) showed significant increases of 0.96% and 1.77% oIoU than 'Enhanced LE' method on each dataset. These interesting results demonstrated the effectiveness of the visual expression *itself*. In addition, we compared our full method with the all-pixel method (row4) that uses all visual pixels as visual guidance elements. Even though the all-pixel method can provide the unique visual information to the network, our method showed 0.99% and 1.66% higher oIoU on each dataset. This indicates that producing visual expression tokens is better to improve the ability to understand the visual contexts of the target regions than using all of pixels, by selectively exploiting the semantic information relevant to the target regions and considering the contextual information between the visual expression tokens.

In Figure 4, we also displayed the precision-recall curves. The area under curve (AUC-PR) summarizes the overall performance of the model across different threshold values. As shown in Figure 4, 'VE only' method maintained its advantage in precision over the 'LE only' and 'Enhanced LE' methods. Our full model had the highest AUC-PR.

4.3.2 Effectiveness of Adaptive Selection

Selection method. To better select the visual information, we experimented with the thresholding method and the top-k method in Table 4 (a). The top-k selection method showed higher oIoU performance than the thresholding method. This indicates that the top-k selection is more adaptive for select-

Method	mIoU (%)	oIoU (%)
w/o selection	66.77 (-1.86)	64.71 (-1.99)
sigmoid (> 0.5)	67.42 (-1.21)	65.35 (-1.35)
top-k ranked selection	**68.63**	**66.70**

(a) **Selection method**

Global	Local	mIoU (%)	oIoU (%)
✓	✗	66.52 (-2.11)	64.34 (-2.36)
✗	✓	66.65 (-1.98)	64.67 (-2.03)
✓	✓	**68.63**	**66.70**

(b) **Utilization of linguistic cues**

Method	mIoU (%)	oIoU (%)
✗	66.85 (-1.78)	64.80 (-1.90)
✓	**68.63**	**66.70**

(c) **Semantic refinement**

Method	mIoU (%)	oIoU (%)
✗	66.98 (-1.65)	64.88 (-1.82)
✓	**68.63**	**66.70**

(d) **Considering visual relationship**

Method	mIoU (%)	oIoU (%)
✗	67.54 (-1.09)	65.43 (-1.27)
✓	**68.63**	**66.70**

(e) **Supervised by the contrastive loss**

Method	mIoU (%)	oIoU (%)
w/o top-k mask	66.91 (-1.72)	64.95 (-1.75)
w/ top-k mask	**68.63**	**66.70**

(f) **Normalization with top-k mask**

Table 4: Ablation studies for the design of our visual expression extractor on RefCOCO+ *val*. Our default setting is marked in gray. The drops are relative to our default setting.

ing the semantic visual information than the thresholding method that depends on high relevance to the linguistic cues. Even if there is insufficient target information in the linguistic cues, this adaptive selection allows our module to exploit the useful pixel information, which has weak relevance to the linguistic features but is helpful for target segmentation. Thus, our module can prevent over-reliance on the high relevance to linguistic cues during the selection step and enhance the adaptability for the diverse linguistic expressions and image contexts.

Utilization of linguistic cues. We conducted the ablation on the effectiveness of the global-local linguistic cues in the adaptive selection step. In Table 4 (b), compared to our full model, removing the use of the local linguistic cues showed a 2.36% drop in oIoU. In addition, removing the use of the global linguistic cue showed a 2.03% drop in oIoU. These results indicate that using both global and local linguistic cues allows the visual expression tokens to consider the enriched visual contextual information of the fine-grained target regions, as each linguistic cue has different contextual information.

4.3.3 Effectiveness of Semantic Refinement

In Table 4 (c), we conducted the ablation on the refinement step with the selected pixels. This result highlights that the refinement step, which enables the aggregated visual tokens to dynamically capture the semantic information from the selected information, is effective than simply aggregating the selected information for producing more semantic visual expression tokens.

4.3.4 Effect of Visual Relationship Modeling

In Table 4 (d), we conducted the ablation on the visual relationship modeling step. This result indicates that each token of the visual expression acquires the visual context information for target regions by considering the relationship between each visual token. Therefore, the visual expression tokens can improve the ability to the visual understanding of the target regions.

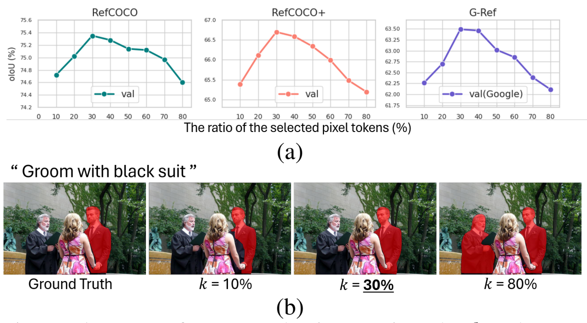

Figure 5: (a) Performance by increasing the k value on three splits. (b) Segmentation results at different k.

4.3.5 Design Choices

Supervision by the contrastive loss. In Table 4 (e), we experimented on supervising the relevance score map by the pixel contrastive loss (Eq.3). This result indicates that the contrastive loss helps to monitor the selection of the useful pixel tokens associated with the correct target region and to prevent the high relevance scores between the linguistic features and incorrect regions.

Normalization with top-k mask. We ablated on applying a softmax normalization with the top-k mask to the relevance scores (Eq.6). In Table 4 (f), normalizing without the top-k mask showed a significant performance drop. This means that using the selected pixels relevant to the target regions is beneficial for robust segmentation than using all pixels including the irrelevant pixels.

4.3.6 Analysis on Number of k

We experimented on the value of k, which is the ratio of the pixel tokens selected for the visual expression extraction to adaptively exploit the useful visual information. Compared to the k values of 10 and 80, the k of 30 showed higher oIoU in Figure 5(a). In addition, as shown in Figure 5(b), the k of 30 segmented more clearly, while the k of 10 missed some part of the target regions and the k of 80 even segmented other object regions. The smaller number of k resulted in a lack of information, where the useful visual information cannot be sufficiently exploited. In contrast, the larger number of k resulted in including the noise information

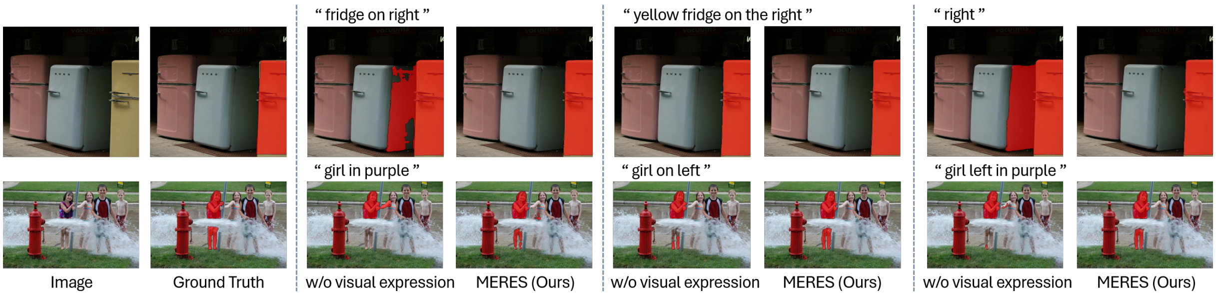

Figure 6: Visualization of our method and ablated model on various language expressions describing the same target object in the image. Additional results are in Appendix F.

Figure 7: Visualization of our method and the previous state-of-the-art methods [15, 62] on the different types of the images and language expressions. Additional results are in Appendix F.

and degrades the guidance capability. Therefore, the optimal k can adaptively select the semantic visual information and filter out noise components to improve the robustness of the guidance capacity.

4.4 Qualitative Results

Comparison to the ablation model. In addition to the comparison in Figure 2, we compared the segmentation results for different language expressions describing the same target in Figure 6. Our method consistently predicted the accurate regions by leveraging the visual expression, which complements the linguistic guidance information, while the ablated model predicted inconsistently and segmented the incorrect regions. These results indicate that our method enhances the adaptability to various language expressions and image inputs for robust segmentation, and improves the ability to comprehend visual contexts of the target regions.

Comparison to the state-of-the-arts. In Figure 7, we compared with previous methods, which use only the enhanced linguistic tokens as the guidance set, on diverse types of inputs. Our method segmented more clearly for the complex and ambiguous language expressions (*e.g.*, (a) and (b)) and the complicated images (*e.g.*, (c) and (d)), whereas other methods incorrectly predicted and uncertainly segmented the regions. These results indicate that our approach is more effective in improving visual understanding of the target regions. More cases, including longer expressions, are in Appendix F.

Method	Venue	LLM	Vision Encoder	RefCOCO			RefCOCO+			G-Ref	
				val	test A	test B	val	test A	test B	$val_{(U)}$	$test_{(U)}$
LISA-7B [38]	CVPR '24	✓	SAM-H [33]	74.1	76.5	71.1	62.4	67.4	56.5	66.4	68.5
PixelLM [57]	CVPR '24	✓	CLIP-VIT-L	73.0	76.5	68.2	66.3	71.7	58.3	**69.3**	**70.5**
MERES	-	✗	Swin-B	**75.4**	**78.0**	**71.9**	**66.7**	**72.1**	**59.9**	65.8	66.9

Table 5: Comparison with LLM-based RES models.

Figure 8: Qualitative comparison to a LLM-based RES model on RefCOCO+. More results are in Appendix G.

5 Conclusion

We proposed a novel Multi-Expression guidance framework for Referring Expression Segmentation (MERES), which enables visual expression tokens as well as linguistic expression tokens to be used as the guidance elements, to complement the linguistic guidance capacity by effectively providing visual contexts of the target regions. To produce semantic visual expression, we design a visual expression extractor that adaptively selects the useful visual information related to target regions from image contexts and allows the visual expression tokens to consider the richer visual contexts. This enhances the adaptability to diverse image and language inputs, and improves visual understanding of the fine-grained target regions. Extensive experiments demonstrated the effectiveness of our approach on three public referring expression segmentation benchmarks.

6 Limitations

With the development of LLMs [64, 18, 17, 74] for vision-language multi-modal tasks [36, 30, 41, 7], LLM-based RES models [38, 57] have been actively explored in this task. For further exploration, we conducted comparison with these models in Table 5. Compared to LLM-based models, our model showed lower performance on the most challenging dataset, G-Ref, which consists of the difficult language samples. As shown in Figure 15 of the Appendix, the reason for this is that due to the much smaller model parameters and smaller training datasets, our model lacks the reasoning ability for the implicit and detailed descriptions in comparison to the LLM. This finding suggests that our performance bottleneck may still lie in understanding the language expressions on this task, while our model has better performance than the existing state-of-the-art models in Table 1. The possible solutions to overcome this limitation are to train with the large-scale image-text datasets, to exploit the stronger language model (*e.g.*, LLMs), and to leverage the language learning techniques [21, 66, 13, 29, 39, 67, 8, 2, 1, 5].

Another finding is that our model surprisingly showed better performance on RefCOCO and RefCOCO+ in Table 5. This finding indicates that our model has a stronger ability to understand the visual contexts of the target regions than LLM-based models, which rely on the generated linguistic token (*i.e.*, the LLM's ability) for their segmentation ability, as shown in Figure 8. In future work, beyond the reliance on the LLM's ability, the exploration of extending our approach to combine with the LLM has the potential for broader impact and further generalization.

References

[1] Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*.

[2] Laurestine Bradford, Timothy John O'Donnell, and Siva Reddy. 2024. A compositional typed semantics for universal dependencies. *arXiv preprint arXiv:2403.01187*.

[3] Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. 2019. See-through-text grouping for referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7454–7463.

[4] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

[5] Xuanda Chen, Timothy O'Donnell, and Siva Reddy. 2024. When does word order matter and when doesn't it? *arXiv preprint arXiv:2402.18838*.

[6] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can pre-trained vision and language models answer visual information-seeking questions? *arXiv preprint arXiv:2302.11713*.

[7] Yangyi Chen, Xingyao Wang, Manling Li, Derek Hoiem, and Heng Ji. 2023. Vistruct: Visual structural knowledge extraction via curriculum guided code-vision representation. *arXiv preprint arXiv:2311.13258*.

[8] Yulin Chen, Ning Ding, Xiaobin Wang, Shengding Hu, Hai-Tao Zheng, Zhiyuan Liu, and Pengjun Xie. 2023. Exploring lottery prompts for pre-trained language models. *arXiv preprint arXiv:2305.19500*.

[9] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. 2021. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5559–5568.

[10] Zesen Cheng, Kehan Li, Peng Jin, Siheng Li, Xiangyang Ji, Li Yuan, Chang Liu, and Jie Chen. 2024. Parallel vertex diffusion for unified visual grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1326–1334.

[11] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

[12] Jean-Baptiste Cordonnier, Aravindh Mahendran, Alexey Dosovitskiy, Dirk Weissenborn, Jakob Uszkoreit, and Thomas Unterthiner. 2021. Differentiable patch selection for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2351–2360.

[13] Ganqu Cui, Wentao Li, Ning Ding, Longtao Huang, Zhiyuan Liu, and Maosong Sun. 2022. Decoder tuning: Efficient language understanding as decoding. *arXiv preprint arXiv:2212.08408*.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

[16] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. 2022. Davit: Dual attention vision transformers. In *European conference on computer vision*, pages 74–92. Springer.

[17] Ning Ding, Yulin Chen, Ganqu Cui, Xingtai Lv, Ruobing Xie, Bowen Zhou, Zhiyuan Liu, and Maosong Sun. 2024. Mastering text, code and math simultaneously via fusing highly specialized language models. *arXiv preprint arXiv:2403.08281*.

[18] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.

[19] Zihan Ding, Zi-han Ding, Tianrui Hui, Junshi Huang, Xiaoming Wei, Xiaolin Wei, and Si Liu. 2022. Ppmn: Pixel-phrase matching network for one-stage panoptic narrative grounding. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5537–5546.

[20] Seungju Han, Junhyeok Kim, Jack Hessel, Liwei Jiang, Jiwan Chung, Yejin Son, Yejin Choi, and Youngjae Yu. 2023. Reading books is great, but not if you are driving! visually grounded reasoning about defeasible commonsense norms. *arXiv preprint arXiv:2310.10418*.

[21] Arian Hosseini, Siva Reddy, Dzmitry Bahdanau, R Devon Hjelm, Alessandro Sordoni, and Aaron Courville. 2021. Understanding by understanding not: Modeling negation in language models. *arXiv preprint arXiv:2105.03519*.

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

[23] Y Hu, Q Wang, W Shao, E Xie, Z Li, J Han, and P Luo. 2023. Beyond one-to-one: rethinking the referring image segmentation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV) Proceedings*. Institute of Electrical and Electronics Engineers (IEEE).

[24] Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. 2020. Bi-directional relationship inferring network for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4424–4433.

[25] Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. 2020. Referring image segmentation via cross-modal progressive comprehension. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10488–10497.

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

[27] Zhenchao Jin, Xiaowei Hu, Lingting Zhu, Luchuan Song, Li Yuan, and Lequan Yu. 2024. Idrnet: Intervention-driven relation network for semantic segmentation. *Advances in Neural Information Processing Systems*, 36.

[28] Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, and Tieniu Tan. 2021. Locate then segment: A strong pipeline for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9858–9867.

[29] Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. Text encoders bottleneck compositionality in contrastive vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[30] Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's" up" with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*.

[31] Beoungwoo Kang, Seunghun Moon, Yubin Cho, Hyunwoo Yu, and Suk-Ju Kang. 2024. Metaseg: Metaformer-based global contexts-aware network for efficient semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 434–443.

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

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

[34] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

[35] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

[36] Benno Krojer, Vaibhav Adlakha, Vibhav Vineet, Yash Goyal, Edoardo Ponti, and Siva Reddy. 2022. Image retrieval from contextual descriptions. *arXiv preprint arXiv:2203.15867*.

[37] Julia Kruk, Caleb Ziems, and Diyi Yang. 2023. Impressions: Visual semiotics and aesthetic impact understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12273–12291.

[38] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2023. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*.

[39] Jaewook Lee, Seongsik Park, Seong-heum Park, Hongjin Kim, and Harksoo Kim. 2023. A framework for vision-language warm-up tasks in multimodal dialogue models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2789–2799.

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

[41] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

[42] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

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

[44] Chang Liu, Henghui Ding, Yulun Zhang, and Xudong Jiang. 2023. Multi-modal mutual attention and iterative interaction for referring image segmentation. *IEEE Transactions on Image Processing*.

[45] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. 2017. Recurrent multimodal interaction for referring image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1271–1280.

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

[47] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. 2022. Ts2-net: Token shift and selection transformer for text-video retrieval. In *European Conference on Computer Vision*, pages 319–335. Springer.

[48] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.

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

[50] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. 2020. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 10034–10043.

[51] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

[52] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. 2016. 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*, pages 792–807. Springer.

[53] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

[54] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

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

[56] Abdullah Rashwan, Jiageng Zhang, Ali Taalimi, Fan Yang, Xingyi Zhou, Chaochao Yan, Liang-Chieh Chen, and Yeqing Li. 2024. Maskconver: Revisiting pure convolution model for panoptic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 851–861.

[57] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. 2023. Pixellm: Pixel reasoning with large multimodal model. *arXiv preprint arXiv:2312.02228*.

[58] Hongje Seong, Seoung Wug Oh, Joon-Young Lee, Seongwon Lee, Suhyeon Lee, and Euntai Kim. 2021. Hierarchical memory matching network for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12889–12898.

[59] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. 2018. Key-word-aware network for referring expression image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 38–54.

[60] Jae-hun Shim, Hyunwoo Yu, Kyeongbo Kong, and Suk-Ju Kang. 2023. Feedformer: revisiting transformer decoder for efficient semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2263–2271.

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

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

[63] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

[64] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

[65] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

[66] Dong Wang, Ning Ding, Piji Li, and Hai-Tao Zheng. 2021. Cline: Contrastive learning with semantic negative examples for natural language understanding. *arXiv preprint arXiv:2107.00440*.

[67] Hao Wang, Xiahua Chen, Rui Wang, and Chenhui Chu. 2023. Vision-enhanced semantic entity recognition in document images via visually-asymmetric consistency learning. *arXiv preprint arXiv:2310.14785*.

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

[69] Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2023. Uniir: Training and benchmarking universal multimodal information retrievers. *arXiv preprint arXiv:2311.17136*.

[70] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. 2024. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. *arXiv preprint arXiv:2401.13560*.

[71] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. 2022. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165.

[72] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. 2023. Semantics-aware dynamic localization and refinement for referring image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3222–3230.

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

[74] Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, et al. 2024. Advancing llm reasoning generalists with preference trees. *arXiv preprint arXiv:2404.02078*.

[75] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. 2021. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402.

[76] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2023. Opt: Open pre-trained transformer language models, 2022. *URL https://arxiv. org/abs/2205.01068*, 3:19–0.

[77] Zicheng Zhang, Yi Zhu, Jianzhuang Liu, Xiaodan Liang, and Wei Ke. 2022. Coupalign: Coupling word-pixel with sentence-mask alignments for referring image segmentation. *Advances in Neural Information Processing Systems*, 35:14729–14742.

[78] Daquan Zhou, Qibin Hou, Linjie Yang, Xiaojie Jin, and Jiashi Feng. 2022. Token selection is a simple booster for vision transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[79] Qinfeng Zhu, Yuanzhi Cai, Yuan Fang, Yihan Yang, Cheng Chen, Lei Fan, and Anh Nguyen. 2024. Samba: Semantic segmentation of remotely sensed images with state space model. *arXiv preprint arXiv:2404.01705*.

[80] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. 2023. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127.

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

Appendix

- Code and README file were submitted as a zip file for reproducibility.
- In Appendix A, we present additional explanations to clarify the differences between our approach and previous approaches.
- In Appendix B, we provide the performance comparison with other methods that are trained with the additional large scale text-image pair datasets.
- In Appendix C, we provide the additional implementation details.
- In Appendix D, we provide the additional details for datasets.
- In Appendix E, we provide the additional details for the generalization setting.
- In Appendix F, we provide additional qualitative results on the various types of language expressions, the different language expressions describing the same target object, and the long and difficult language expressions.
- In Appendix G, we present the additional qualitative comparison with the LLM-based RES model (*i.e.*, LISA [38]).

A Difference from Previous Approaches

In Figure 9, we illustrated the guidance sets of previous approaches and our approach to clarify the differences. Previous approaches used the various linguistic guidance elements to guide the network to the target regions. As shown in (a), CRIS [68] used the linguistic encoder features as the elements of the guidance set. As shown in (b), VG-LAW [61] used the layer-specific linguistic features as the guidance elements, which embedded for each layer of the vision encoder. As shown in (c), BRINet [24], ReSTR [32], VLT [15], DMMI [23], and CGFormer [62] used the visual-attended linguistic features as the elements of the guidance set, which are enhanced by referring to the vision features; we called these features as the enhanced linguistic expression tokens in this paper. As shown in (d), JMCELN [26] and ReLA [43] used the dynamic multi-modal tokens, which dynamically capture the region and language features by using the learnable tokens, as the guidance elements. As shown in (e),

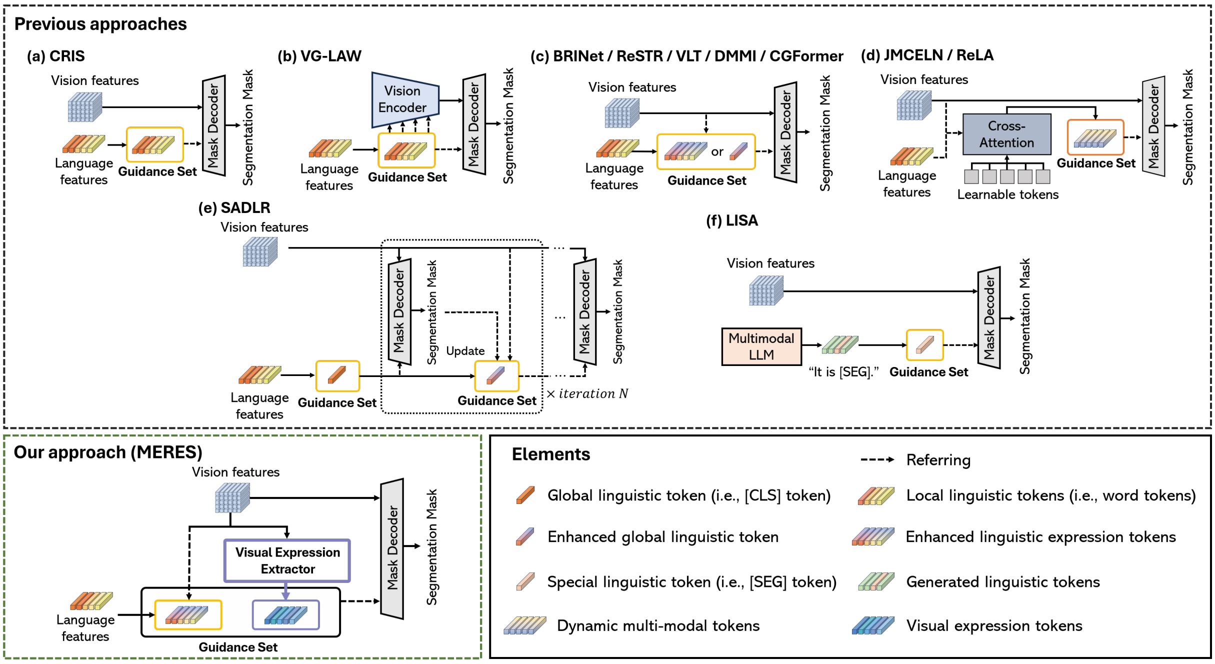

Figure 9: Guidance set comparison of previous approaches (*i.e.*, CRIS [68], VG-LAW [61], BRINet [24], ReSTR [32], VLT [15], DMMI [23], CGFormer [62], JMCELN [26], ReLA [43], SADLR [72] and LISA [38]) and our approach. Previous approaches leverage various guidance sets to guide the network to the target regions. Different from these approaches, our approach enables the visual expression tokens as well as the enhanced linguistic expression tokens to be used as the elements of the guidance set, to complement the linguistic guidance capacity by effectively providing the visual understanding of the fine-grained target regions.

SADLR [72] used the global linguistic features as the guidance elements, which are iteratively updated with the pooled visual vector based on the previous iteration's prediction mask. As shown in (f), LISA [38], the LLM-based RES model, used the special linguistic token (*i.e.*, `[SEG]` token) generated by the multimodal LLM as the guidance elements.

Different from these approaches, our approach uses not only the enhanced linguistic expression tokens but also the visual expression tokens as the elements of the guidance set, as illustrated in Figure 9. Our visual expression tokens complement the linguistic guidance capacity by effectively providing the visual contexts of the fine-grained target regions. Therefore, our method allows the network to avoid relying on the linguistic guidance.

B Comparison with Other RES Methods

To further analysis of our method, we compared our model with other RES methods [80, 81, 46] that use the additional large scale text-image pair datasets [54, 34, 4] at training. PolyFormer [46] showed higher performance on four splits (*i.e.*, RefCOCO+ val. & test A, and G-Ref val$_{(U)}$ & test$_{(U)}$). However, despite the unfair condition of not using any additional large-scale text-image datasets at training, our model outperformed it on the other splits. These results demonstrate the great adaptability of our approach.

C Additional Implementation Details

Experimental Settings. Our method was implemented in PyTorch [53]. We used the AdamW [49] optimizer with initial learning rate of 3e-5 and adopted the polynomial learning rate decay scheduler. The input image resolution was 480×480.

Evaluation Metrics. Following previous works, we adopted the overall intersection-over-union (oIoU), mean intersection-over-union (mIoU), and precision at 0.5, 0.7 and 0.9 thresholds. The oIoU is the ratio between the total intersection regions and the total union regions of all test samples. The mIoU is the average of IoUs between the predicted mask and the ground truth of all test samples. The precision is the percentage of test samples that have an IoU score higher than a threshold.

D Additional Details for Datasets

RefCOCO & RefCOCO+. These two datasets are distributed under the Apache-2.0 license, and are collected from the two-player game [73]. The

Method	Venue	Vision Encoder	RefCOCO			RefCOCO+			G-Ref		
			val	test A	test B	val	test A	test B	$\text{val}_{(U)}$	$\text{test}_{(U)}$	$\text{val}_{(G)}$
X-Decoder (B) [80]	CVPR '23	DaViT-B [16]	-	-	-	-	-	-	64.5	-	-
SEEM (B) [81]	NeurIPS '23	DaViT-B	-	-	-	-	-	-	65.0	-	-
PolyFormer [46]	CVPR '23	Swin-B	74.82	76.64	71.06	**67.64**	**72.89**	59.33	**67.76**	**69.05**	-
MERES (Ours)	-	Swin-B	**75.35**	**77.97**	**71.94**	66.70	72.08	**59.85**	65.78	66.93	**63.49**

Table 6: oIoU performance comparison with other RES models, which use the additional large scale text-image pair datasets at training, on three public referring expression segmentation benchmarks. (U): UMD split. (G): Google split. The best score is in **bold**.

Dataset	Split	Max	Min	Mean
RefCOCO	train	39	1	3.5
	val	21		3.6
	test A	23		3.4
	test B	27		3.6
RefCOCO+	train	24	1	3.5
	val	22		3.6
	test A	16		3.3
	test B	22		3.8
G-Ref	train	46	1	8.5
	$\text{val}_{(U)}$	37		8.5
	$\text{test}_{(U)}$	32		8.4
	$\text{val}_{(G)}$	37		8.5

Table 7: Length of the language expression samples on each split of all datasets.

evaluation sets of RefCOCO and RefCOCO+ are splitted into the validation subset, the test A subset and the test B subset. The images of the testA subset contain the multiple people, and the images of the testB subset contain the multiple instances of all other objects. RefCOCO+, which forbids the words about the absolute locations in the language expressions, is more challenging than RefCOCO.

G-Ref. This dataset is distributed under the CC-BY 4.0 license, and is collected on Amazon Mechanical Turk. We use both UMD [52] and Google [51] partitions for the evaluation. The UMD partition splits the evaluation set into the validation subset and the test subset. The Google partition consists of only the validation set. The average length of the language expressions is 8.4 words. This means that the G-Ref dataset contains longer and more complex language expressions than the RefCOCO and RefCOCO+ datasets. Thus, G-Ref is the most challenging dataset.

In Table 7, we present the length of the language expression samples on each split of three datasets.

E Additional Details for Generalization Setting

To further validate the generalization ability of our model, we experimented on the generalization setting introduced by [62]. These setting splits the RES datasets into the seen and unseen categories on MSCOCO [42] of the open-vocabulary detection [75]. The training set contains only the seen categories, and the test set consists of the seen subset and the unseen subset. Following the previous work [62], we adopted the text encoder of CLIP [55] as the language encoder for a fair comparison in this experiment, and trained our model for 50 epochs.

F Additional Qualitative Results

In addition to the comparison (Figure 7) with the previous methods [15, 62] that use the visual-attended linguistic tokens as the elements of the guidance set, we compared with the other state-of-the-art method [43], which uses the dynamic multi-modal tokens as the elements of the guidance set, on longer and more complex language expressions in Figure 10. These results demonstrate that our MERES can effectively enhance the adaptability to the various cases and improve visual understanding of the fine-grained target regions than other approaches.

In Figure 11, we visualized more results of our MERES and the previous methods [15, 62], which use the visual-attended linguistic tokens as the elements of the guidance set, on the challenging types of language expressions to verify the robustness of our method, such as typos and slang, which make it difficult for the network to refer to the linguistic contexts. Compared to previous methods [15, 62], our MERES correctly determined the target regions. In Figure 12, we visualized additional qualitative results on various types of the language expressions and the images to clearly demonstrate the high level of competence in understanding the

Figure 10: Additional qualitative comparison of our method and the existing state-of-the-art method (*i.e.*, ReLA [43]), which uses the dynamic multi-modal tokens as the elements of the guidance set, on longer and more difficult language expressions.

Figure 11: Visualization of our method and the previous methods [15, 62], which use the visual-attended linguistic tokens as the elements of the guidance set, on the challenging types of the linguistic expressions such as typos and slang.

context of the target regions. Our MERES showed more accurate segmented regions than the previous state-of-the-art methods for the diverse expressions describing the relative location (*e.g.* "animal behind fence" and "banana closest to apples"), color (*e.g.* "white" and "beige") and other attributes (*e.g.*, "200999", "empty" and "with handles").

In Figure 13, we visualized additional results of our full model and the ablation model for two or three different language expressions describing the same object. Our method showed robust segmentation for various language expressions, whereas the ablation model segmented the non-target regions or did not highlight the target regions.

G Additional Qualitative Comparison with LISA

In Figure 14, we present the additional comparison on RefCOCO+ with LISA [38], which leverages the capabilities of the Large Language Model (LLM). Our model showed robust segmentation for the challenging target regions (*e.g.*, "Guy sitting with black shirt" and "The person a white shirt and white hat"). These results indicate that our approach is more effective in understanding the visual contexts of the target object compared to the LLM-based method. On the other hand, for the challenging language expressions that are too difficult even for humans to understand, our model showed failure cases as shown in Figure 15, while LISA correctly detected the target object. This means that the LLM-based method has a great ability to comprehend the meaning of the implicit and complex language expressions. Therefore, in future work, the exploration of combining our approach's strength with the LLM's strength has the potential for broader impact on this task.

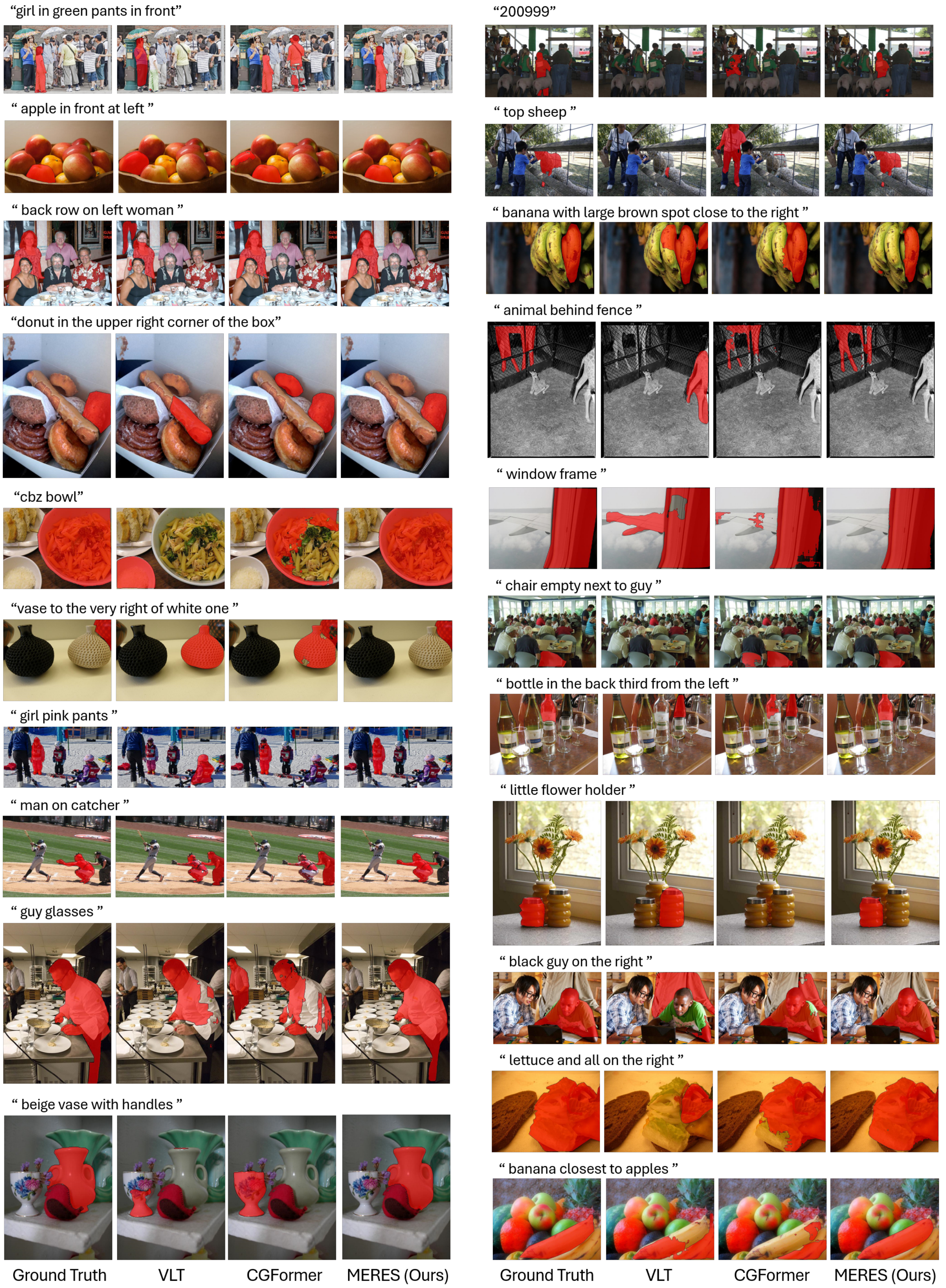

Figure 12: Additional qualitative comparison of the proposed method and the previous state-of-the-art methods on more diverse language expressions and images.

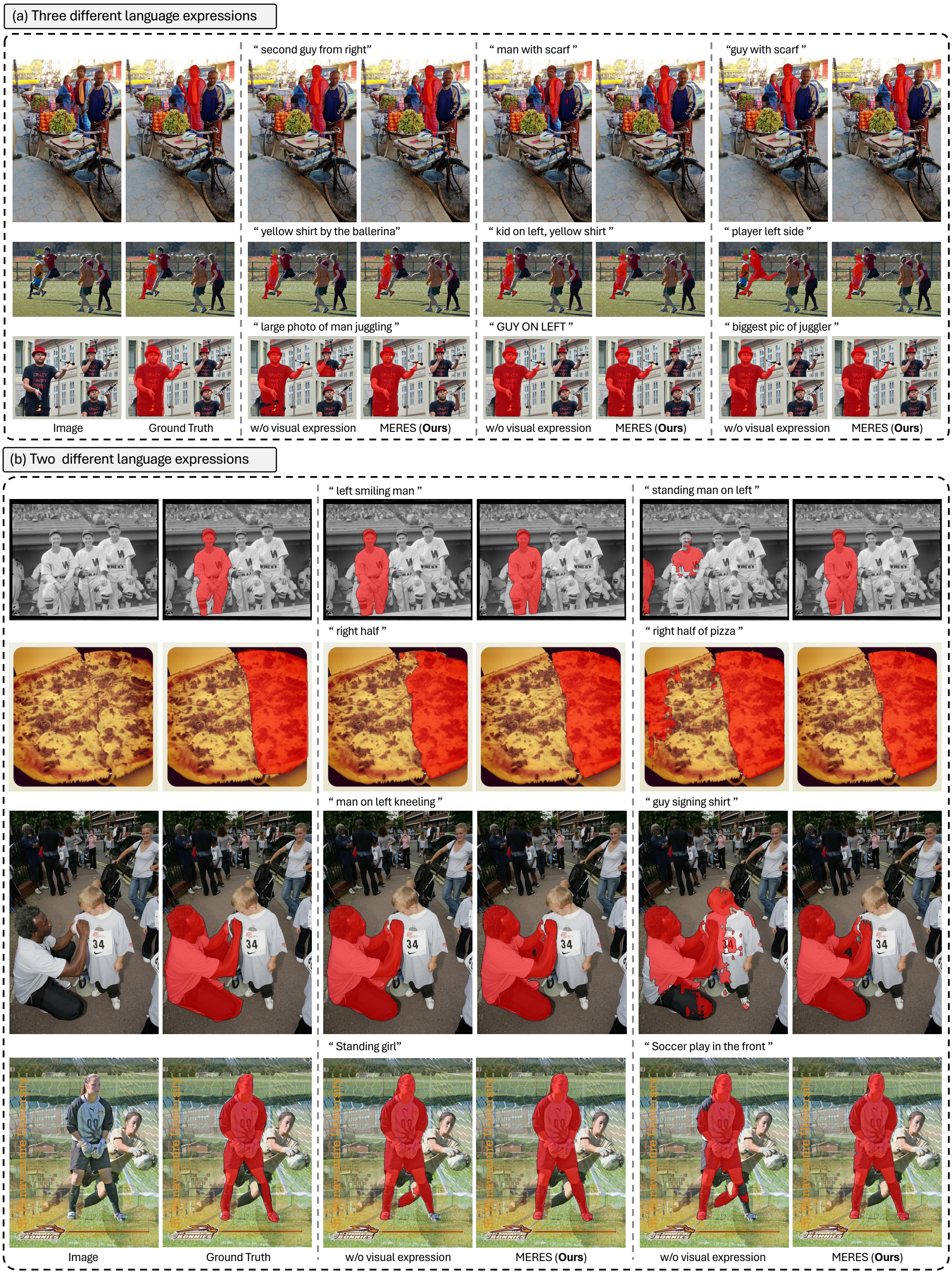

Figure 13: Additional qualitative comparison of the proposed method and the ablated model on different language expressions describing the same object in the image.

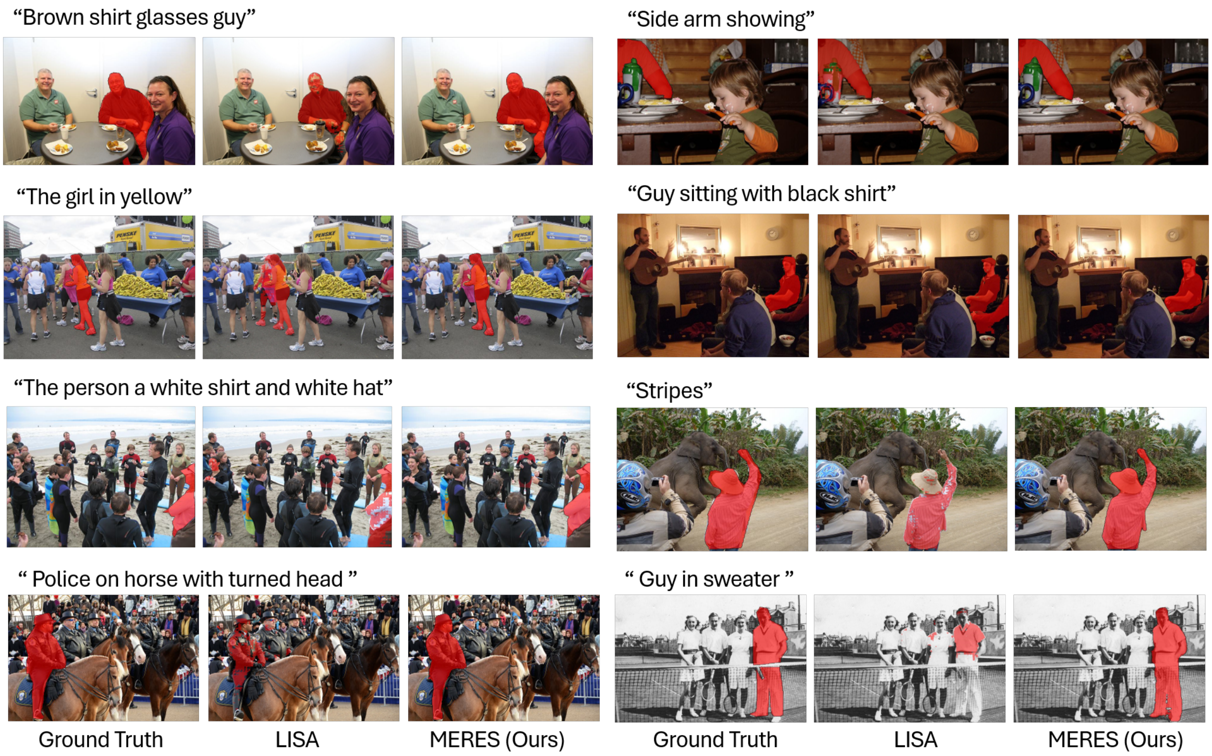

Figure 14: Additional qualitative results of our MERES and LISA [38] on RefCOCO+ dataset.

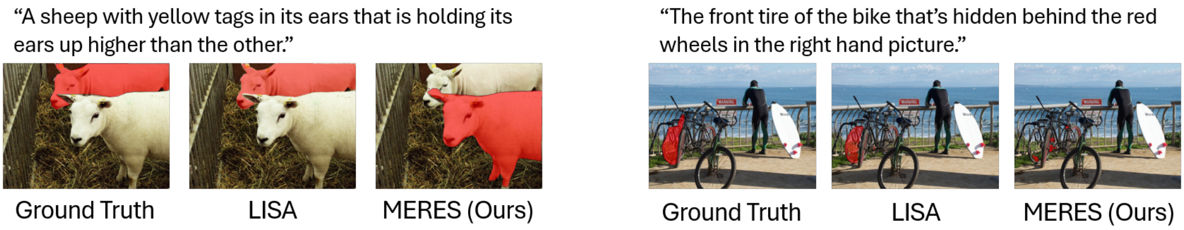

Figure 15: Failure cases of our MERES on G-Ref dataset.