Progressive Cross-Modal Grounding with Residual Token Selection for Referring Image Segmentation

Anonymous ACL submission

Abstract

Referring Image Segmentation (RIS) aims to predict pixel-level masks for objects described by natural language expressions. Existing RIS methods typically rely on static one-stage vision-language fusion, which often leads to insufficient cross-modal alignment under complex referring expressions. We argue that effective referring image grounding should progressively evolve through semantic filtering, hierarchical cross-modal interaction, and visually guided semantic refinement. To this end, we propose a Progressive Vision-Language Grounding framework (PVLG-RIS) that progressively refines multimodal representations for dense vision-language grounding. Specifically, the framework first suppresses redundant visual tokens while preserving global contextual semantics through residual token aggregation. It then progressively aligns visual and textual representations across multiple Transformer layers, enabling gradual refinement from coarse localization to fine-grained grounding. Furthermore, global visual context is exploited to adaptively refine textual semantics and enhance language-guided visual representation learning. Finally, multi-scale visual representations are aggregated through feature pyramid fusion and decoded using a prototype-based mask generation strategy. Extensive experiments on RefCOCO, RefCOCO+, and RefCOCOg demonstrate that PVLG-RIS consistently outperforms existing methods, particularly under long and ambiguous referring expressions. Further analysis shows that progressive grounding produces more robust and interpretable vision-language alignment for RIS.

1 Introduction

1.1 Problem Definition

RIS (Ding et al., 2021; Hu et al., 2016; Yao et al., 2021) aims to ground a natural language expression onto pixel-level visual regions within an image. Given an image $I \in R^{H \times W \times 3}$ and a referring expression $T = \{w_1, w_2, \ldots, w_N\}$, the objective is to predict a binary mask $M \in R^{H \times W}$ corresponding to the target entity described by the expression.

Unlike standard semantic segmentation, RIS requires cross-modal reasoning over attributes, spatial relations, and contextual dependencies. The model must align textual semantics with hierarchical visual representations while resolving ambiguity in complex referring expressions.

1.2 Core Challenge

Despite recent progress, RIS remains challenging due to three key issues. First, referring expressions are often long and compositional, requiring fine-grained alignment between words and visual regions (Yu et al., 2018; Hui et al., 2020). Second, Vision Transformer and CLIP-based encoders (Ouyang et al., 2023; Yue et al., 2024) generate dense patch tokens, many of which are irrelevant to the target expression, introducing redundancy that degrades localization quality. Third, most existing methods (Yang et al., 2021; Ding et al., 2021; Kim et al., 2022) perform either weak or late-stage fusion between vision and language, which limits early interaction and results in suboptimal region-text alignment for dense prediction.

1.3 Motivation

Recent CLIP-based RIS methods (Wang et al., 2021; Luo et al., 2021; Yao et al., 2021) show strong global alignment ability, but still struggle with spatially precise grounding. We observe three limitations in existing pipelines. First, treating all visual tokens equally introduces redundant features that are not relevant to the referring expression (Ouyang et al., 2023). Second, single-scale fusion fails to capture both global context and local boundary details (Yang et al., 2021). Third, visual features are often weakly conditioned on language during intermediate representation learning, lim-

iting fine-grained alignment. These observations suggest that RIS benefits from explicit token filtering, multi-scale feature integration, and stronger language-conditioned modulation.

1.4 Contributions

In this work, we propose a Progressive Vision-Language Grounding framework for referring image segmentation (PVLG-RIS). Unlike static one-stage fusion methods (Shang et al., 2024; Huang et al., 2020), our framework progressively refines cross-modal alignment through semantic filtering, hierarchical interaction, and language refinement.

- **Progressive Cross-Modal Alignment Formulation.** We reformulate RIS as a progressive vision-language alignment problem rather than conventional one-stage multimodal fusion. We argue that effective grounding should evolve through hierarchical semantic filtering, cross-modal interaction, and language refinement, enabling more robust alignment between textual semantics and visual regions.

- **Progressive Vision-Language Grounding Framework.** We propose a unified progressive grounding framework that performs: (1) language-guided visual token filtering to suppress redundant visual regions, (2) hierarchical cross-modal interaction across Transformer layers for progressive semantic alignment, and (3) visually guided language refinement for adaptive language-conditioned feature modulation. Unlike independent fusion modules, our framework enables dynamic hierarchical multimodal reasoning for dense vision-language grounding.

- **Comprehensive Evaluation and Analysis.** Extensive experiments on RefCOCO, RefCOCO+, and RefCOCOg show consistent improvements, especially under long and ambiguous referring expressions. Ablation studies and qualitative results further verify the effectiveness of progressive grounding and adaptive token refinement for robust multimodal alignment.

2 Related Work

2.1 Referring Image Segmentation

RIS aims to localize and segment target objects in an image according to natural language expressions. Existing fully supervised approaches achieve strong performance by learning from pixel-level annotations paired with referring expressions. However, these methods rely heavily on expensive dense annotations and often struggle to generalize under complex semantic descriptions and open-world scenarios.To alleviate annotation dependency and improve generalization, recent studies explored weakly supervised, generalized, and training-free RIS frameworks. GRES (Liu et al., 2023) investigated generalized referring segmentation, while recent methods further improved semantic grounding through enhanced spatial reasoning, alignment-aware learning, and large-scale vision-language pretraining (Shi and Yang, 2023; Suo et al., 2023; Chen et al., 2026). Large language model assisted grounding strategies (Qi et al., 2024) were also explored to enhance semantic understanding and open-vocabulary grounding capability. In addition, recent reasoning-oriented RIS frameworks such as DeRIS (Dai et al., 2025) emphasized iterative multimodal reasoning and perception-cognition interaction for more robust segmentation.

2.2 Vision-Language Grounding

RIS requires accurate alignment between visual regions and natural language expressions. Early RIS methods such as MAttNet (Yu et al., 2018), CMSA (Ye et al., 2019), RMI (Liu et al., 2017), RRN (Li et al., 2018), and keyword-aware frameworks (Shi et al., 2018) mainly relied on recurrent language encoders and convolutional backbones for multimodal interaction. Dynamic interaction strategies were further explored to improve adaptive vision-language alignment (Margffoy-Tuay et al., 2018; Yang et al., 2019), while pipeline-based methods such as LTS (Jing et al., 2021) decomposed RIS into sequential localization and segmentation stages.

Recent transformer-based approaches significantly improved long-range dependency modeling and cross-modal reasoning. LAVT (Yang et al., 2021), VLTVG (Ding et al., 2021), ReSTR (Kim et al., 2022), and LQMFormer (Shah et al., 2024) employed transformer architectures for dense multimodal interaction. CLIP-based methods such as CRIS (Wang et al., 2021) further enhanced semantic grounding through pretrained vision-language representations. Recent studies also explored hierarchical reasoning and alignment-aware interaction for improved grounding (Dai et al., 2025; Chen et al., 2026). However, most existing RIS

frameworks still rely on direct or single-stage cross-modal fusion, limiting progressive semantic refinement under complex referring expressions. Beyond fully supervised settings, generalized, zero-shot, and weakly supervised RIS have also been investigated. GRES (Liu et al., 2023) explored generalized referring segmentation, while spatial-aware methods (Shi and Yang, 2023; Suo et al., 2023) improved zero-shot grounding through enhanced semantic reasoning. Large language model assisted grounding (Qi et al., 2024) and weakly supervised pseudo-labeling strategies (Wu et al., 2023) were further explored to reduce annotation dependency and improve open-vocabulary grounding capability.

Nevertheless, effectively integrating token-level semantic filtering, hierarchical grounding, and progressive language refinement within a unified vision-language grounding framework remains underexplored.

3 Method

3.1 Overview

Given an image I and a referring expression T, RIS aims to predict a pixel-level mask corresponding to the target object described in the text. As illustrated in Fig. 1, we formulate RIS as a Progressive Vision-Language Grounding (PVLG-RIS) process, where multimodal representations are progressively refined from noisy token space to semantically grounded object-level alignment.

Semantic Filtering. We first apply a Dynamic Residual Token Selection (DRTS) operator to suppress redundant visual tokens based on language-aware importance estimation, while preserving global context via residual aggregation. This yields a compact, noise-reduced visual representation for subsequent alignment.

Hierarchical Cross-Modal Alignment. Next, a Progressive Cross-Modal Grounding (PCMG) operator performs layer-wise interaction between visual and textual features. Rather than a single fusion step, it gradually strengthens alignment from coarse spatial correspondence to fine-grained grounding.

Visual-Guided Semantic Refinement. Finally, a Visual-Guided Language Refinement (VGLR) operator updates textual embeddings under global visual context and modulates visual features in a language-conditioned manner. This bidirectional refinement further enhances discriminability for ambiguous expressions.

Through these progressive transformations, representations evolve from raw multimodal features into tightly aligned and semantically grounded embeddings, which are aggregated via an FPN-style fusion module and decoded into segmentation masks (Feng et al., 2021).

3.2 Dynamic Residual Token Selection

As illustrated in Fig. 2, Vision Transformers produce dense visual tokens that often contain background regions irrelevant to the referring expression. These redundant tokens may interfere with cross-modal alignment and degrade localization performance. To address this issue, we propose a **Dynamic Residual Token Selection (DRTS)** module that explicitly separates language-relevant tokens from redundant ones while preserving global contextual information.

Given visual tokens $V = \{v_i\}_{i=1}^{N-1}$ and language tokens $T = \{t_j\}_{j=1}^{L}$, we estimate the semantic relevance of each visual token by computing its averaged similarity to all language tokens:

$$s_i = \frac{1}{L} \sum_{j=1}^{L} v_i^\top t_j \tag{1}$$

where s_i denotes the relevance score of the i-th visual token with respect to the referring expression. By averaging similarities across all text tokens, the score captures the overall semantic consistency between a visual region and the complete language query.

Based on the relevance scores, we retain the top-K visual tokens as the language-relevant token subset V_{sel}:

$$V_{\text{sel}} = \text{TopK}(V, s) \tag{2}$$

The remaining low-relevance tokens are aggregated into a compact residual contextual token rather than being directly discarded. Let $\mathcal{R}$ denote the index set of non-selected tokens. We first compute normalized contextual weights:

$$\alpha_i = \frac{\exp(s_i)}{\sum_{k \in \mathcal{R}} \exp(s_k)} \tag{3}$$

The residual token is then constructed by weighted aggregation:

$$v_r = \sum_{i \in \mathcal{R}} \alpha_i v_i \tag{4}$$

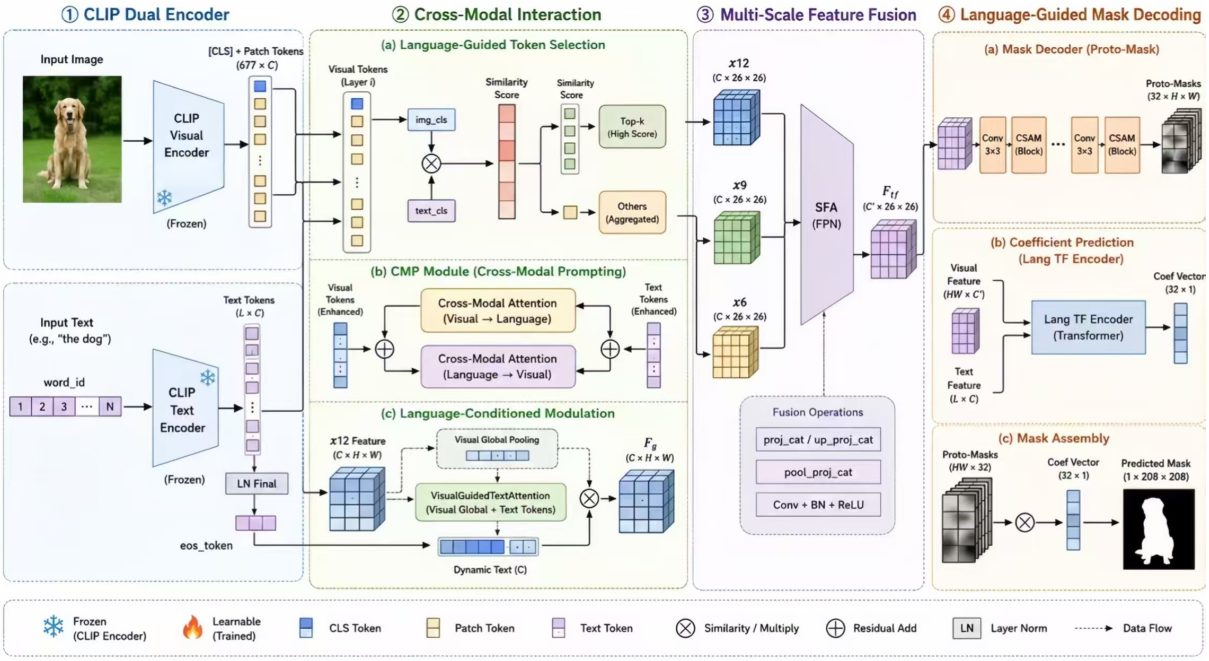


Figure 1: Overview of the proposed Progressive Vision-Language Grounding framework for robust referring image segmentation(PVLG-RIS). The model consists of a CLIP dual encoder, progressive cross-modal interaction modules, multi-scale feature fusion, and a language-guided prototype mask decoder.

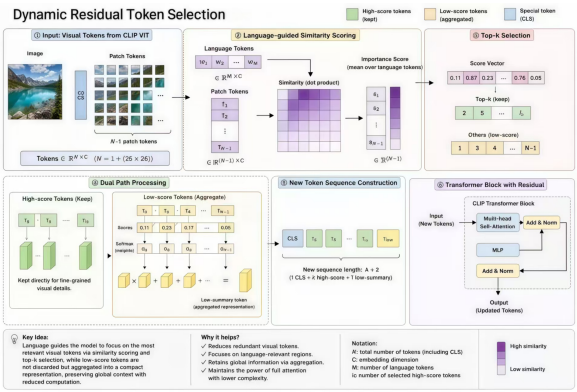


Figure 2: Overview of the proposed Dynamic Residual Token Selection (DRTS) module.

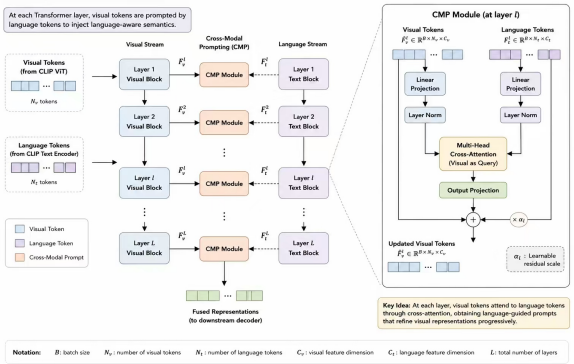


Figure 3: Overview of the proposed Progressive Cross-Modal Grounding Framework (PCMG) module.

Here, v_r serves as a compact contextual representation that summarizes suppressed background regions and non-local visual dependencies. Unlike hard token pruning, residual aggregation preserves global contextual awareness while reducing redundant token interaction.

The final compact token sequence is formed by concatenating the CLS token, selected tokens V_{sel}, and the residual token v_r. This compact representation is processed by a Transformer block for feature refinement, where the residual token interacts with both selected visual tokens and the CLS token through self-attention. After refinement, the updated compact features are projected back to the original token space via residual broadcasting.

Compared with direct token pruning strategies, DRTS suppresses irrelevant visual regions while maintaining global contextual information through residual aggregation, thereby improving cross-modal alignment under complex referring expressions.

3.3 Progressive Cross-Modal Grounding Framework

To enhance fine-grained vision-language interaction, we propose a Progressive Cross-Modal

Grounding (PCMG) module, as illustrated in Fig. 3. Instead of one-shot multimodal fusion, PCMG progressively injects linguistic semantics into visual representations across multiple Transformer layers, enabling hierarchical cross-modal alignment from coarse localization to fine-grained semantic grounding.

Early Transformer layers mainly capture coarse object localization and spatial correspondence, while deeper layers progressively refine attribute-level semantics and discriminative object details. Since visual representations are iteratively refined through language-aware interaction, subsequent layers operate on increasingly discriminative multimodal features.

Given visual tokens $F_v^l \in R^{B\times N_v\times C_v}$ and language tokens $F_t^l \in R^{B\times N_t\times C_t}$ at the l-th Transformer stage, we first project them into a shared embedding space:

$$\hat{F}_v^l = \mathrm{LN}(W_v F_v^l), \qquad \hat{F}_t^l = \mathrm{LN}(W_t F_t^l) \tag{5}$$

where W_v and W_t are learnable projection layers and $\mathrm{LN}(\cdot)$ denotes Layer Normalization.

We then perform language-guided cross-attention, where visual tokens dynamically retrieve expression-relevant semantic cues from language tokens:

$$T_{v\leftarrow t}^l = \mathrm{MHA}\left(\hat{F}_v^l, \hat{F}_t^l, \hat{F}_t^l\right) \tag{6}$$

where $\mathrm{MHA}(\cdot)$ denotes multi-head attention.

The attended features are fused back into the visual representations through residual refinement:

$$F_v^{l+1} = F_v^l + \alpha_l \cdot W_o\left(T_{v\leftarrow t}^l\right) \tag{7}$$

where W_o is a learnable projection layer and α_l is a learnable residual scaling factor controlling semantic injection strength.

Through repeated cross-modal interaction, PCMG progressively conditions visual representations on language semantics, enabling multimodal alignment to evolve from coarse spatial grounding to discriminative object-level reasoning.

3.4 Visual-Guided Language Refinement

Existing language modulation strategies in RIS often rely on static textual embeddings, overlooking the compositional and context-dependent nature of referring expressions under complex attributes, spatial relations, and ambiguous semantics. To address this issue, we propose a **Visual-Guided Language Refinement (VGLR)** module that formulates language guidance as a visually grounded token-level semantic refinement process. Instead of directly using fixed textual embeddings for visual modulation, VGLR dynamically updates language semantics under global visual context before multimodal fusion, enabling adaptive semantic disambiguation and finer vision-language alignment.

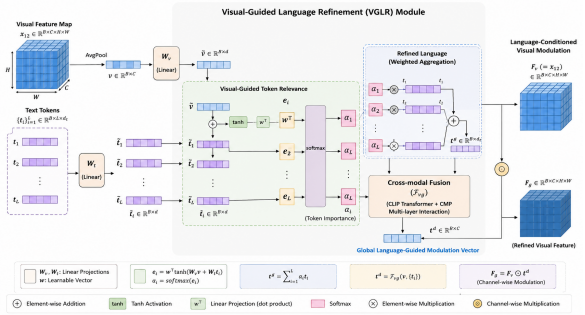


Figure 4: Visual-Guided Language Refinement module.

As illustrated in Fig. 4, VGLR first extracts global visual semantics to refine language tokens and subsequently uses the refined language representations to modulate visual features. Different from PCMG, which progressively injects language semantics into visual representations, VGLR mainly focuses on refining linguistic semantics under scene-level visual context.

Given a visual feature map $x_{12} \in R^{B\times C\times H\times W}$, we first obtain a compact global visual representation:

$$v = \mathrm{AvgPool}(x_{12}) \in R^{B\times C} \tag{8}$$

For each language token $\{t_i\}_{i=1}^L$, we estimate its visual-conditioned semantic relevance:

$$e_i = w^\top \tanh(W_v v + W_t t_i), \quad \alpha_i = \mathrm{softmax}(e_i) \tag{9}$$

where α_i denotes the contextual importance of the i-th token. This token reweighting emphasizes informative semantics, such as attributes and spatial relations, while suppressing irrelevant linguistic components.

The language representations are further refined through a Transformer-based refinement block:

$$t^d = \mathcal{F}_{\mathrm{vg}}(v, \{t_i\}) \tag{10}$$

where $\mathcal{F}_{\mathrm{vg}}$ denotes a visual-guided cross-modal Transformer layer built upon the shared CLIP embedding space. Unlike conventional gating strategies, $\mathcal{F}_{\mathrm{vg}}$ explicitly performs token-level semantic

Method	Visual	Language	RefCOCO			RefCOCO+			G-Ref			Avg
			val	testA	testB	val	testA	testB	val(U)	test(U)	val(G)	
CGAN (Luo et al., 2020)	DarkNet-53	Bi-GRU	64.86	68.04	62.07	51.03	55.51	44.06	51.01	51.69	46.54	54.98
LAVT (Yang et al., 2021)	Swin-B	BERT	72.73	75.82	68.79	62.14	68.38	55.10	61.24	62.09	60.50	65.20
VLT (Ding et al., 2021)	Swin-B	BERT	72.96	75.96	69.60	63.53	68.43	56.92	63.49	66.22	62.80	66.66
SLViT (Ouyang et al., 2023)	SegNeXt	BERT	74.02	76.91	70.62	64.07	69.28	56.14	62.75	63.57	60.94	66.48
LQMFormer (Shah et al., 2024)	Swin-B	BERT	74.16	76.82	71.04	65.91	71.84	57.59	64.73	66.04	62.97	67.90
SADLR (Yang et al., 2023)	Swin-B	BERT	74.24	76.25	70.06	64.28	69.09	55.19	63.60	63.56	61.16	66.38
DMMI (Hu et al., 2023)	Swin-B	BERT	74.13	77.13	70.16	63.98	69.73	57.03	63.46	64.19	61.98	66.87
MCRES (Xu et al., 2023a)	Swin-B	BERT	74.92	76.98	70.84	64.32	69.68	56.64	63.51	64.90	61.63	67.05
ReLA (Liu et al., 2023)	Swin-B	BERT	73.82	76.48	70.18	66.04	71.02	57.65	65.00	65.97	62.70	67.65
SGEPG (Wu et al., 2023)	ResNet-101	BiLSTM	69.61	72.34	65.67	55.21	60.67	46.76	61.33	60.61	–	61.52
CrossVLT (Cho et al., 2023)	Swin-B	CLIP	73.44	76.16	70.15	63.60	69.10	55.23	62.28	63.75	–	66.76
CRIS (Wang et al., 2021)	CLIP-R101	CLIP	70.47	73.18	66.10	62.27	68.08	53.68	59.87	60.36	–	64.25
OmniSegNet (Zheng et al., 2025)	Swin-B	CLIP	74.02	76.93	70.63	65.95	71.30	58.31	64.90	66.27	–	68.54
STCM (Zhang et al., 2026)	Swin-B	CLIP	66.62	68.49	57.86	58.79	62.70	49.25	–	–	53.96	59.67
CLIP-based Methods												
CM-MaskSD (Wang et al., 2023)	CLIP-ViT-B	CLIP	72.18	75.21	67.91	64.47	69.29	56.55	62.67	62.69	–	66.37
RISCLIP-B (Kim et al., 2024)	CLIP-ViT-B	CLIP	73.57	76.46	69.76	65.53	70.61	55.49	64.10	65.09	–	67.58
ETRIS (Xu et al., 2023b)	CLIP-ViT-B	CLIP	70.51	73.51	66.63	60.10	66.89	50.17	59.82	59.91	57.88	62.82
ASDA (Yue et al., 2024)	CLIP-ViT-B	CLIP	75.06	77.14	71.36	66.84	71.13	57.83	65.73	66.45	63.55	68.34
Ours	CLIP-ViT-B	CLIP	**75.30**	**77.29**	**71.49**	**68.02**	**72.76**	**59.16**	**66.49**	**66.95**	**65.58**	**69.22**

Table 1: Comparison with state-of-the-art referring image segmentation methods on RefCOCO, RefCOCO+, and G-Ref. All results are reported using the corresponding visual–language backbones. **U** and **G** denote the UMD and Google splits of G-Ref, respectively. "–" indicates unavailable results. IoU (%) is used as the evaluation metric.

refinement conditioned on visual context, improving the discriminability of ambiguous referring expressions.

Finally, the refined language representation is used to modulate visual features:

$$F_g = F_v \odot t^d \tag{11}$$

where $t^d \in R^{B\times C}$ is broadcast along the spatial dimension as a language-guided reweighting vector. Since modulation is conditioned on visually refined language semantics rather than static textual embeddings, the resulting multimodal representations become more robust for long and complex referring expressions.

Overall, VGLR establishes a sequential visual-to-language refinement and language-to-visual modulation process, strengthening compositional vision-language grounding for RIS.

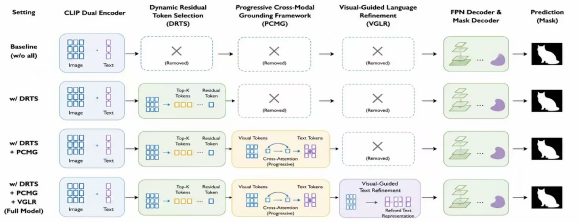
4 Experiments

4.1 Dataset and Metrics

We evaluate our method on three referring expression grounding benchmarks: RefCOCO, RefCOCO+, and RefCOCOg (Wu et al., 2020), all built on MS COCO. RefCOCO contains 19,994 images and 142,209 expressions for 50,000 objects, featuring short descriptions with spatial relations and simple attributes. RefCOCO+ follows a similar setting but removes explicit spatial words, making grounding more appearance- and context-dependent, with 141,564 expressions over 49,856 objects. RefCOCOg contains longer and more natural expressions (avg. 8.4 words), with 104,560 expressions over 54,822 objects in 26,711 images, requiring stronger language understanding and multi-step reasoning. Following prior work, we evaluate model performance using two standard metrics: overall Intersection-over-Union (IoU) and Precision@X. IoU measures the pixel-wise overlap between predicted and ground-truth masks, serving as the primary metric for segmentation quality. Precision@X reports the percentage of test samples whose IoU exceeds threshold X, where $X \in \{0.5, 0.7, 0.9\}$. While Precision $X \in \{0.5\}$ reflects coarse localization accuracy, higher thresholds provide stricter evaluation of fine-grained mask quality. All metrics are computed on the official evaluation splits for fair comparison with existing methods.

4.2 Implementation Details

We build our model on CLIP ViT-B/16 with image resolution 416×416 and a maximum text length of 17. Experiments are conducted on RefCOCO, RefCOCO+, and RefCOCOg under DDP with 2 GPUs. We optimize using Adam with an initial learning rate of 5×10^{-5}, applying lower rates ($0.1\times$) to CLIP encoders. The model is trained for 32 epochs with batch size 16, and the learning rate is decayed at epochs 18 and 28. We use mixed-precision training with loss weights $\lambda_{seg} = 0.1$ and $\lambda_{coord} = 5$. The segmentation threshold is 0.35. The model is initialized from pretrained CLIP ViT-



Setting	RefCOCO+(val)				RefCOCO+(testA)				RefCOCO+(testB)			
	P@0.5 ↑	P@0.7 ↑	P@0.9 ↑	IoU ↑	P@0.5 ↑	P@0.7 ↑	P@0.9 ↑	IoU ↑	P@0.5 ↑	P@0.7 ↑	P@0.9 ↑	IoU ↑
Baseline (w/o all)	74.97	67.54	25.99	66.84	81.52	73.80	26.38	71.13	64.02	53.97	22.86	57.83
w/ DRTS	75.72	67.42	26.22	66.51	82.58	75.25	27.29	71.62	64.75	55.49	23.39	58.33
w/ DRTS + PCMG	76.64	68.92	25.30	67.20	82.64	74.99	26.51	71.68	65.35	55.55	23.33	58.36
Full Model (w/ DRTS + PCMG + VGLR)	**77.28**	**69.15**	**26.80**	**68.02**	**83.54**	**76.21**	**26.91**	**72.76**	**65.96**	**55.61**	**23.52**	**59.16**

Figure 5: Ablation studies on the RefCOCO+ dataset.

B/16.

4.3 Quantitative Results

We evaluate our method on RefCOCO, RefCOCO+, and RefCOCOg under standard protocols, and compare it with both non-CLIP and CLIP-based RIS methods.

As shown in Table 1, non-CLIP methods such as CGAN (Luo et al., 2020), LAVT (Yang et al., 2021), VLT (Ding et al., 2021), and SLViT (Ouyang et al., 2023) achieve competitive performance but are limited by weaker vision-language alignment. Transformer-based approaches such as LQMFormer (Shah et al., 2024), SADLR (Yang et al., 2023), and MCRES (Xu et al., 2023a) further improve cross-modal reasoning, achieving stronger segmentation accuracy. Our method consistently outperforms all baselines across all datasets. In particular, it achieves 69.22% overall performance, surpassing ASDA by a clear margin, especially on challenging subsets.

4.4 Ablation Study

To evaluate the effectiveness of each component, we conduct ablation studies on the RefCOCO+ dataset under standard validation and test splits.

As shown in Table 5, the baseline model achieves relatively lower performance due to the lack of effective cross-modal interaction and fine-grained token modeling. Introducing the Dynamic Residual Token Selection (DRTS) module brings consistent improvements by removing redundant visual tokens and enhancing language-relevant feature selection. Furthermore, the Progressive Cross-Modal Grounding (PCMG) module further improves performance by enabling hierarchical interaction between visual and textual representations across Transformer layers, leading to better cross-modal alignment. Finally, the Visual-Guided Language Refinement (VGLR) module enhances language representations by incorporating global visual context, producing more discriminative textual embeddings for segmentation. When all components are combined, the full model achieves the best performance on RefCOCO+, demonstrating that each module contributes complementary improvements to vision-language grounding.

4.5 Efficiency Analysis

We analyze the computational efficiency of the proposed PVLG-RIS framework in terms of parameter complexity, inference throughput, and training configuration. The full model contains approximately 219.0M parameters, compared with 193.1M parameters in the ASDA baseline. We further evaluate the inference throughput of the proposed framework. Under a batch size of 16 and an input resolution of 416×416, the model processes one batch in approximately 0.137 seconds during inference on NVIDIA GPUs, corresponding to an average throughput of roughly 117 FPS (images per second) under batched evaluation. These results indicate that the proposed progressive grounding strategy introduces limited additional computational overhead while maintaining efficient inference performance.

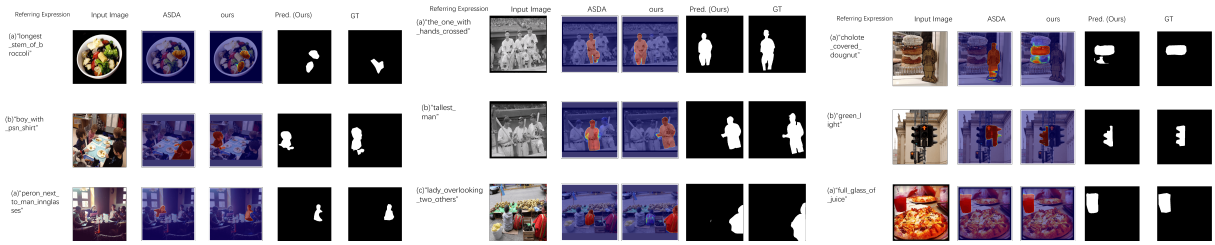
4.6 Visualization

We compare our method with the recent CLIP-based approach ASDA (Yue et al., 2024) on the RefCOCO+ validation and test splits. As shown in Fig. 6, we visualize the referring expressions, grounding heatmaps, predicted masks, and ground-truth annotations. Compared with ASDA, our method generates more accurate and concentrated attention regions, especially under complex expressions involving attribute reasoning, spatial relations, and cluttered backgrounds. The predicted masks produced by our framework exhibit clearer object boundaries and better semantic consistency with the textual query.

4.7 Discussion and Analysis

4.7.1 Why Progressive Prompting Works

The proposed Progressive Cross-Modal Grounding (PCMG) module progressively injects linguistic semantics into visual representations across multiple Transformer layers, enabling hierarchical vision-language interaction. Unlike one-stage fusion strategies, progressive prompting allows the model to refine cross-modal alignment gradually from low-level spatial cues to high-level semantic



(a) Val (b) TestA (c) TestB

Figure 6: Qualitative comparison between ASDA and our method on RefCOCO+ splits. From left to right: referring expression, input image, ASDA heatmap, our heatmap, predicted mask, and ground-truth annotation.

understanding. We observe that this design produces more concentrated grounding responses and improves localization consistency under complex referring expressions. In particular, progressive interaction helps the model better capture attribute descriptions, spatial relations, and small target objects. The improvements indicate that hierarchical cross-modal refinement is more effective than direct feature fusion for dense vision-language grounding.

4.7.2 Effect of Residual Token Compression

Dense visual tokens generated by Vision Transformers often contain substantial redundant background information, which may interfere with language-aware grounding. The proposed Dynamic Residual Token Selection (DRTS) module addresses this issue by dynamically preserving informative visual tokens while compactly aggregating redundant ones into residual representations.

Visualization results show that DRTS suppresses irrelevant background regions and encourages the model to focus on language-related object areas. Compared with direct token pruning, residual token compression maintains global contextual information while reducing noisy feature interaction. This mechanism improves both computational efficiency and grounding robustness, particularly in cluttered scenes and complex visual environments.

5 Limitations

Although the proposed framework achieves strong performance across multiple RIS benchmarks, several limitations remain.

First, the proposed framework still faces challenges under highly ambiguous referring expressions or severe occlusion scenarios, where discriminative visual cues are insufficient for reliable grounding. Second, the progressive cross-modal interaction strategy introduces additional computational overhead compared with lightweight one-stage fusion methods, especially when multiple Transformer layers are involved. Third, the current framework is mainly designed for static image referring segmentation and has not been fully validated in video-based, temporal grounding, or open-vocabulary segmentation settings. Finally, while the proposed modules demonstrate complementary improvements, the framework still relies on large-scale pretrained vision-language backbones, which may limit deployment in resource-constrained environments.

6 Conclusion

In this paper, we propose a Progressive Vision-Language Grounding framework for referring image segmentation (PVLG-RIS). Built upon a pretrained CLIP dual-encoder, our framework progressively enhances cross-modal alignment through three complementary components: Dynamic Residual Token Selection (DRTS), Progressive Cross-Modal Grounding (PCMG), and Visual-Guided Language Refinement (VGLR). Instead of one-stage multimodal fusion, our method progressively refines visual-language representations through hierarchical token interaction and language-guided modulation. Specifically, DRTS suppresses redundant visual tokens while preserving language-relevant context through residual aggregation. PCMG injects linguistic semantics into visual representations across Transformer layers for hierarchical cross-modal reasoning. VGLR further refines textual semantics under global visual guidance, enabling more discriminative language-conditioned modulation for dense prediction.

Overall, our work highlights the importance of progressive cross-modal interaction and adaptive token refinement for vision-language grounding, providing insights into multimodal reasoning and dense vision-language understanding.

References

Tongfei Chen, Shuo Yang, Yuguang Yang, Linlin Yang, Runtang Guo, Changbai Li, He Long, Chunyu Xie, Dawei Leng, and Baochang Zhang. 2026. Amlris: Alignment-aware masked learning for referring image segmentation. *arXiv preprint arXiv:2602.22740*.

Yubin Cho, Hyunwoo Yu, and Suk-Ju Kang. 2023. Cross-aware early fusion with stage-divided vision and language transformer encoders for referring image segmentation. *IEEE Transactions on Multimedia*, 26:5823–5833.

Ming Dai, Wenxuan Cheng, Jiang-jiang Liu, Sen Yang, Wenxiao Cai, Yanpeng Sun, and Wankou Yang. 2025. Deris: Decoupling perception and cognition for enhanced referring image segmentation through loop-back synergy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19936–19946.

Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. 2021. Vision-language transformer and query generation for referring segmentation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16301–16310.

Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. 2021. Encoder fusion network with co-attention embedding for referring image segmentation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15501–15510.

Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. 2016. Segmentation from natural language expressions. *ArXiv*, abs/1603.06180.

Yutao Hu, Qixiong Wang, Wenqi Shao, Enze Xie, Zhenguo Li, Jungong Han, and Ping Luo. 2023. Beyond one-to-one: Rethinking the referring image segmentation. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4044–4054.

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. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10485–10494.

Tianrui Hui, Si Liu, Shaofei Huang, Guanbin Li, Sansi Yu, Faxi Zhang, and Jizhong Han. 2020. Linguistic structure guided context modeling for referring image segmentation. *ArXiv*, abs/2010.00515.

Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, and Tieniu Tan. 2021. Locate then segment: A strong pipeline for referring image segmentation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9853–9862.

Nam-Won Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. 2022. Restr: Convolution-free referring image segmentation using transformers. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18124–18133.

Seoyeon Kim, Minguk Kang, Dongwon Kim, Jaesik Park, and Suha Kwak. 2024. Extending CLIP's image-text alignment to referring image segmentation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4611–4628, Mexico City, Mexico. Association for Computational Linguistics.

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

Chang Liu, Henghui Ding, and Xudong Jiang. 2023. Gres: Generalized referring expression segmentation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23592–23601.

Chenxi Liu, Zhe L. Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Loddon Yuille. 2017. Recurrent multimodal interaction for referring image segmentation. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1280–1289.

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

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2021. Clip4clip: An empirical study of clip for end to end video clip retrieval. *Neurocomputing*, 508:293–304.

Edgar Margffoy-Tuay, Juan Pérez, Emilio Botero, and Pablo Arbeláez. 2018. Dynamic multimodal instance segmentation guided by natural language queries. In *European Conference on Computer Vision*.

Shuyi Ouyang, Hongyi Wang, Shiao Xie, Ziwei Niu, Ruofeng Tong, Yen-Wei Chen, and Lanfen Lin. 2023. Slvit: Scale-wise language-guided vision transformer for referring image segmentation. In *International Joint Conference on Artificial Intelligence*.

Lu Qi, Yi-Wen Chen, Lehan Yang, Tiancheng Shen, Xiangtai Li, Weidong Guo, Yu Xu, and Ming-Hsuan Yang. 2024. Generalizable entity grounding via assistance of large language model. *arXiv preprint arXiv:2402.02555*.

Nisarg A Shah, Vibashan VS, and Vishal M Patel. 2024. Lqmformer: Language-aware query mask transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12903–12913.

Chao Shang, Zichen Song, Heqian Qiu, Lanxiao Wang, Fanman Meng, and Hongliang Li. 2024. Prompt-driven referring image segmentation with instance contrasting. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4124–4134.

Cheng Shi and Sibei Yang. 2023. Spatial and visual perspective-taking via view rotation and relation reasoning for embodied reference understanding. In *European Conference on Computer Vision*.

Hengcan Shi, Hongliang Li, Fanman Meng, and Q. Wu. 2018. Key-word-aware network for referring expression image segmentation. In *European Conference on Computer Vision*.

Yucheng Suo, Linchao Zhu, and Yi Yang. 2023. Text augmented spatial aware zero-shot referring image segmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1032–1043.

Wenxuan Wang, Jing Liu, Xingjian He, Yisi Zhang, Cheng Chen, Jiachen Shen, Yan Zhang, and Jiangyun Li. 2023. Cm-masksd: Cross-modality masked self-distillation for referring image segmentation. *IEEE Transactions on Multimedia*, 26:6906–6916.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yan Guo, Ming Gong, and Tongliang Liu. 2021. Cris: Clip-driven referring image segmentation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11676–11685.

Cantao Wu, Yi Cai, Liuwu Li, and Jiexin Wang. 2023. Scene graph enhanced pseudo-labeling for referring expression comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11978–11990, Singapore. Association for Computational Linguistics.

Chenyun Wu, Zhe Lin, Scott D. Cohen, Trung Bui, and Subhransu Maji. 2020. Phrasecut: Language-based image segmentation in the wild. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10213–10222.

Li Xu, Mark He Huang, Xindi Shang, Zehuan Yuan, Ying Sun, and Jun Liu. 2023a. Meta compositional referring expression segmentation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19478–19487.

Z. Xu, Zhihong Chen, Yong Zhang, Yibing Song, Xiang Wan, and Guanbin Li. 2023b. Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17457–17466.

Shuang Yang, Liang Li, Miao Tang, Yunchao Xu, Yulun Zhang, and Jingdong Liu. 2023. Semantics-aware dynamic localization and refinement for referring image segmentation. In *AAAI Conference on Artificial Intelligence*.

Sibei Yang, Guanbin Li, and Yizhou Yu. 2019. Dynamic graph attention for referring expression comprehension. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4643–4652.

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip H. S. Torr. 2021. Lavt: Language-aware vision transformer for referring image segmentation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18134–18144.

Lewei Yao, Runhu Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training. *ArXiv*, abs/2111.07783.

Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. 2019. Cross-modal self-attention network for referring image segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10494–10503.

Licheng Yu, Zhe L. Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. 2018. Mattnet: Modular attention network for referring expression comprehension. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1307–1315.

Pengfei Yue, Jianghang Lin, Shengchuan Zhang, Jie Hu, Yilin Lu, Hongwei Niu, Haixin Ding, Yan Zhang, Guannan Jiang, Liujuan Cao, and Rongrong Ji. 2024. Adaptive selection based referring image segmentation. *Proceedings of the 32nd ACM International Conference on Multimedia*.

Congwei Zhang, Zhibin Quan, and Wankou Yang. 2026. Cross-modal semantic token alignment via contrastive learning for weakly-supervised referring image segmentation. *Expert Systems with Applications*, 319:132029.

Qiancheng Zheng, Yunhang Shen, Gen Luo, Baiyang Song, Xing Sun, Xiaoshuai Sun, Yiyi Zhou, and Rongrong Ji. 2025. Omni-referring image segmentation. *arXiv preprint arXiv:2512.06862*.