Resolving the Inherent Contextual Insufficiency in Referring Image Segmentation with Global Semantic Priors

Chong Yi, Jialei Chen, Seigo Ito, Hiroshi Murase, and Daisuke Deguchi

Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi, Japan

Abstract. Referring Image Segmentation methods have made remarkable progress in fine-grained localization but frequently falter in complex scenarios populated by multiple candidates sharing similar visual attributes, often resulting in the erroneous segmentation of distractors. We attribute this limitation to the inherent insufficiency of global context in sparse referring expressions: these brief queries often fail to capture the holistic scene semantics, thereby failing to provide the discriminative cues necessary for precise grounding. To address this issue, we propose the Caption-Guided Attention Network (CGA-Net), a novel framework that leverages generated image captions as a Global Semantic Prior to supplement the sparse linguistic query with structured scene descriptions, thereby strictly strengthening visual–linguistic alignment. Specifically, we introduce a Dual-Path Context Injection Module that simultaneously infuses the global prior into both visual and linguistic modalities, effectively bridging the informational gap between the concise expression and the complex image. Furthermore, we introduce a Semantic Consistency Constraint that aligns the visual representation with the caption to ensure high-level semantic coherence. Extensive evaluations on the RefCOCO, RefCOCO+, and RefCOCOg benchmarks validate that CGA-Net establishes a new state-of-the-art, effectively overcoming the semantic bottleneck in complex scenes.

Keywords: Global Semantic Prior · Vision-Language Models · Semantic Consistency · Referring Image Segmentation

1 Introduction

Referring Image Segmentation (RIS) is a fundamental task in multimodal understanding, aiming to segment specific objects in an image based on natural language expressions. Distinct from conventional semantic segmentation constrained by predefined categories, RIS requires fine-grained semantic alignment between linguistic cues and dense visual content. Early approaches typically adopted modular designs, employing Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) to process textual and visual features independently before fusion [14, 22, 24, 34]. With the advent of Transformer architectures [1, 2, 9], the field has witnessed a paradigm shift towards modeling dense, pixel-level multi-modal interactions. Frameworks like LAVT [33] and

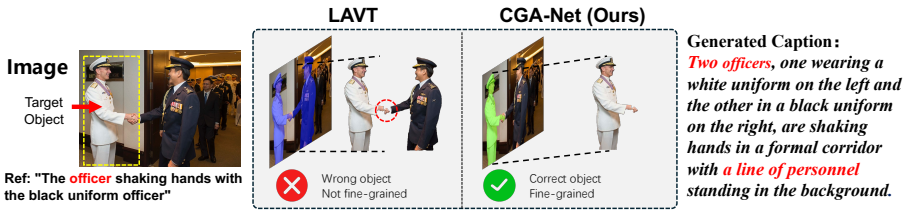

Fig. 1. Visualizing the "Inherent Insufficiency of Global Context". In this scenario populated by multiple officers, the sparse query ("The officer shaking hands...") implies a relationship but lacks the **discriminative cues** to uniquely pinpoint the target among visually similar candidates. Consequently, the baseline falters due to this **informational gap**, over-relying on local attribute matching and missing the holistic scene semantics. In contrast, CGA-Net leverages the **Global Semantic Prior** to **supplement** the query with structured scene descriptions, effectively bridging the gap and correctly localizing the interacting target.

VLT [12] have achieved impressive performance by performing deep fusion at multiple feature stages. More recently, sequence-to-sequence models, such as PolyFormer [20], SeqTR [37] and Prompt-RIS [28], have formulated RIS as a coordinate or patch generation task, achieving significant performance gains.

Despite these architectural advancements, we identify a critical, persistent bottleneck in complex real-world scenarios in sparse referring expressions: the inherent insufficiency of global context. Fig. 1 illustrates this phenomenon. In this scenario, populated by multiple candidates sharing similar visual attributes (e.g., "black uniform"), the sparse instruction "The officer shaking hands..." implies a relationship but fails to capture the holistic scene semantics required to uniquely pinpoint the target. Consequently, even representative frameworks such as LAVT [33] (Fig. 1) falter due to this informational gap. Constrained by the sparse query, they lack the discriminative cues necessary for precise grounding; consequently, they over-rely on local attribute matching (e.g., "uniform") and fail to distinguish the true referent from visually analogous distractors.

We attribute this failure to a common limitation shared by current paradigms. While recent advancements—including prompt-based approaches [28,36], sequence-to-sequence models [20,37], and universal perception frameworks [32]—have improved performance by mining deeper features or introducing learnable prompts, they remain fundamentally constrained by the sparse input itself. These approaches implicitly assume that the provided referring expression contains sufficient information to uniquely distinguish the target if processed with sufficient depth. However, this assumption overlooks the fundamental informational gap between the concise expression and the complex image, rendering internal visual–linguistic alignment insufficient to resolve the semantic bottleneck when the input query inherently lacks discriminative cues.

To address this issue, we propose the Caption-Guided Attention Network (CGA-Net). As illustrated in Fig. 2, previous paradigms have evolved from sim-

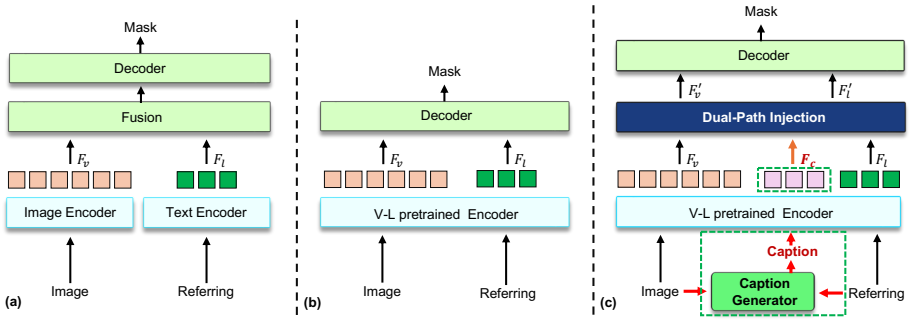

Fig. 2. Comparison of context modeling paradigms. (a) The Sparse-Input Paradigm uses separate encoders for vision and language with late fusion. **(b) The Feature-Augmented Paradigm** advances to pre-trained V-L encoders to enable deep feature interaction. Crucially, both paradigms are limited by the internal information provided by the sparse query. **(c) Our Global Semantic Paradigm** (CGA-Net) fundamentally differs by introducing generated captions as an external semantic source. This explicitly supplements the missing global context, enabling better alignment with the dense visual scene.

ple local matching (Fig. 2(a)) to deep feature interaction (Fig. 2(b)); however, both remain bounded by the insufficient context inherent in the sparse query. To overcome this, we advance the field to a "Global Semantic Paradigm" (Fig. 2(c)). Instead of relying solely on internal alignment, the proposed framework leverages generated image captions as a Global Semantic Prior. The intuition is straightforward: the generated caption captures the holistic scene semantics that the sparse query often lacks. By using this caption to supplement the concise expression with structured scene descriptions, we effectively bridge the informational gap. These additional discriminative cues enable the model to uniquely distinguish the target from visually analogous distractors.

Technically, we design a **Dual-Path Context Injection Module** to integrate the generated caption. This module employs symmetric cross-attention to simultaneously infuse the global prior into both visual and linguistic streams, effectively densifying local features with holistic scene context. Furthermore, a **Semantic Consistency Constraint** is introduced to align the visual representation with the caption to ensure high-level semantic coherence.

The main contributions are summarized as follows:

- **Global Semantic Framework:** CGA-Net is proposed, which integrates a Dual-Path Context Injection Module to infuse generated image captions into local features, effectively compensating for sparse linguistic queries.
- **Semantic Regularization:** A Semantic Consistency Constraint is introduced to explicitly align the visual representation with the global captions. This functions as a regularization term that ensures high-level semantic coherence between modalities during training.

– **Performance:** Extensive experiments on RefCOCO, RefCOCO+, and RefCOCOg benchmarks demonstrate that CGA-Net achieves state-of-the-art performance, effectively resolving the semantic bottleneck and enabling robust referent disambiguation in complex scenes.

2 Related Work

2.1 Referring Image Segmentation

Referring Image Segmentation (RIS) has advanced rapidly from early modular networks to unified Transformer architectures.

Early approaches typically adopted multi-stage pipelines. For instance, MattNet [34] decomposed expressions into subject, location, and relationship modules. MCN [24] introduced a co-attention mechanism to guide visual feature learning, while CMPC [13] proposed a cross-modal progressive comprehension module. With the advent of Transformers, a dominant stream focuses on pixel-level fusion. Following the success of LAVT [33], recent works like VLT [12] and VG-LAW [29] introduce language-adaptive weights to dynamically modulate visual features. Another stream unifies RIS into sequence generation frameworks. SeqTR [37] simplifies RIS as a point prediction task, while PolyFormer [20] outputs polygon coordinates directly. UNINEXT [32] further reformulates RIS as a universal object retrieval task using prompt-guided queries.

Despite their strong performance, these methods generally operate under the assumption that the provided expression is sufficient. As highlighted in GRES [19], standard RIS methods often falter in complex scenarios (e.g., multi-target or ambiguous descriptions) due to the lack of global semantic guidance. Consequently, they suffer from the inherent insufficiency of global context discussed in Section 1, where the sparse query fails to provide the discriminative signals necessary for precise grounding.

2.2 Visual-Language Alignment and Context Modeling

To bridge the gap between vision and language, recent research has explored leveraging external knowledge and advanced context modeling mechanisms.

A prominent direction is leveraging pre-trained Vision-Language Models for alignment. ReferFormer [31] treats language expressions as queries to attend to important regions. CRIS and RIS-CLIP transfer CLIP's knowledge of image-text matching to pixel-level segmentation [4–6, 8, 16, 30]. However, directly mapping global features to dense predictions can be suboptimal. While recent advances in prompt learning [36] and dense adaptation [7, 26] demonstrate that optimizing textual queries or context-aware features [3] significantly improves the adaptation, they primarily focus on optimizing internal alignment. They do not explicitly supplement the input with external scene descriptions. To further enhance semantic understanding, recent studies have introduced mechanisms to capture broader context. VATEX [25] proposes integrating object-centric visual

heatmaps as a prior, while DeRIS [10] proposes a decoupled framework with loop-back synergy to iteratively refine feature understanding. These methods demonstrate the importance of incorporating broader context beyond simple feature matching. Concurrently, the rise of Large Multimodal Models has inspired new paradigms. Models like Grounded-SAM [21] and SEEM [38] combine foundation models for open-set segmentation. LISA [17] further demonstrates that LLMs can perform reasoning segmentation through heavy fine-tuning. Similarly, PixelLM [27] explores efficient pixel-level reasoning with LMMs. In contrast to these heavy frameworks, we explore a more efficient direction by leveraging frozen Generative VLMs (e.g., BLIP-2 [18]) to synthesize a generated image caption as a supplementary knowledge source. Unlike previous approaches that rely on internal feature refinement, CGA-Net introduces captions as a Global Semantic Prior. This supplementary text provides a comprehensive scene summary, aiding the model in effectively resolving the semantic bottleneck with negligible training overhead.

3 Proposed Method

3.1 Preliminary

The goal of Referring Image Segmentation (RIS) is to predict a binary segmentation mask $M \in \{0,1\}^{H\times W}$ that precisely localizes the target object, given an input image $I \in \mathbb{R}^{H\times W\times 3}$ and a natural language referring expression $E = \{w_1, \ldots, w_L\}$, where w denotes referring tokens. Standard approaches typically rely on the local alignment between I and E within an encoder-decoder paradigm. However, as discussed in Section 1, this design is constrained by the inherent insufficiency of global context: the sparse expression E often lacks the holistic scene semantics required to effectively guide the dense visual features, resulting in a significant informational gap.

To bridge this informational gap, we propose the Caption-Guided Attention Network (CGA-Net). It leverages a Global Semantic Prior to supplement the sparse query with structured scene descriptions. As illustrated in Fig. 3, our framework operates through three synergistic stages:

The pipeline commences with the **Base Framework and Feature Extraction** (Section 3.2), where we employ standard backbones to extract multi-scale visual features $\mathcal{V}$ from the image and local linguistic embeddings F_l from the expression. In parallel, to supplement the missing context, we initiate the **Global Semantic Prior Generation** (Section 3.3). This module utilizes a generative Vision-Language Model (VLM) to synthesize a target-aware caption, which is encoded into global semantic features F_c.

Subsequently, these feature streams converge at the core **Dual-Path Context Injection Module** (Section 3.4). Rather than relying on simple concatenation, this module simultaneously infuses the global prior F_c into both visual and linguistic streams via cross-attention, enriching local representations with holistic scene context to bridge the informational gap. Finally, the network predicts the segmentation mask M, supervised by a standard segmentation loss

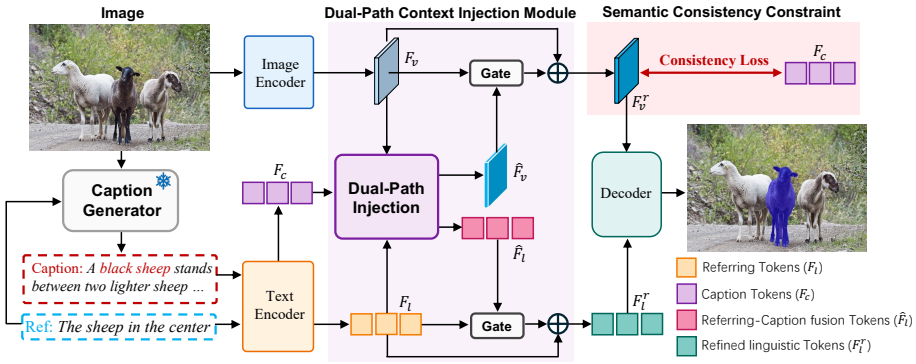

Fig. 3. The overall architecture of the proposed Caption-Guided Attention Network (CGA-Net). Given an input image and a referring expression, we leverage a pre-trained VLM to generate a descriptive caption as a **Global Semantic Prior**. The framework comprises a visual stream, a linguistic stream, and our core **Dual-Path Context Injection Module** that explicitly integrates global semantics into local features. Finally, a **Semantic Consistency Constraint** is applied to enforce high-level alignment between the visual backbone and the global scene context.

and a **Semantic Consistency Constraint** (Section 3.5) that enforces high-level alignment between the visual backbone and the global scene description.

3.2 Base Framework and Feature Extraction

Our CGA-Net is designed as a flexible, model-agnostic module that can be seamlessly integrated into standard Transformer-based RIS architectures (e.g., LAVT [33]). The data flow initiates with the extraction of local visual and linguistic representations from parallel streams.

Visual Stream. Given the input image $I \in \mathbb{R}^{H\times W\times 3}$, we employ a standard hierarchical visual backbone, such as Swin Transformer [23], to extract multi-scale visual features. Let $\mathcal{V} = \{V_i\}_{i=1}^{4}$ denote the feature maps from the four stages of the backbone. To facilitate subsequent multi-modal interaction, we utilize a linear projection layer to map the visual features from the final stage into a common feature dimension D. We denote the resulting flattened visual features as $F_v \in \mathbb{R}^{N_v\times D}$, where N_v is the number of visual tokens. These features encapsulate dense pixel-level information but initially lack explicit awareness of the specific referring target specified by the sparse query.

Linguistic Stream. Simultaneously, the natural language expression E is tokenized and processed by a pre-trained text encoder, such as BERT [11]. This yields a sequence of local linguistic embeddings $F_l \in \mathbb{R}^{L\times D}$, where L is the sequence length and D matches the channel dimension of the visual stream. F_l captures the specific semantic attributes of the expression (e.g., color, category) required for basic matching with local visual features.

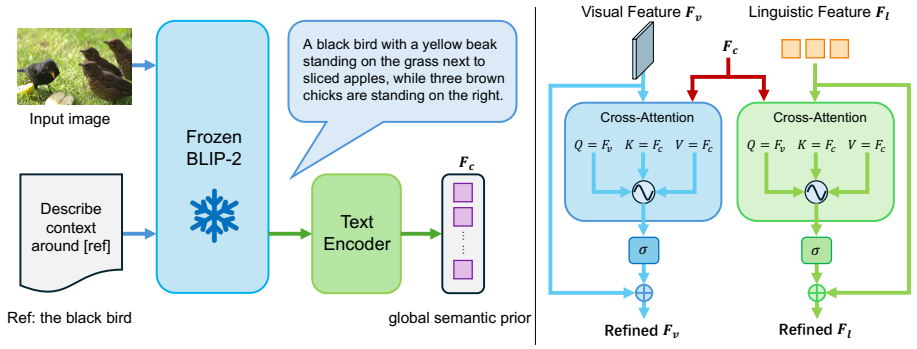

Fig. 4. Illustration of the core components in CGA-Net. (a) **Global Prior Generation:** We construct a target-aware prompt by embedding the referring expression into a query template. The frozen VLM then generates a descriptive caption, which is encoded into the global semantic prior F_c. (b) **Dual-Path Context Injection:** The global prior F_c is injected into both visual and linguistic streams via cross-attention to bridge the semantic gap.

Integration Point. In conventional baselines [33], the extracted F_v and F_l are typically fed directly into a cross-modal decoder to predict the segmentation mask. In our proposed framework, we intercept these feature streams before the decoding stage. We utilize F_v and F_l as the foundational inputs for our **Dual-Path Context Injection Module** (Section 3.4), where they are enriched by the global semantic prior before being passed to the final decoder.

3.3 Global Semantic Prior Generation

To supplement the sparse linguistic input, we leverage the pre-trained BLIP-2 [18] as an external knowledge engine to synthesize a comprehensive scene summary. We keep the VLM parameters frozen to efficiently harness its open-world recognition capabilities without catastrophic forgetting.

Target-Aware Prompting. A generic image caption often focuses on salient objects, potentially missing the specific target described by the referring natural language expression E. To resolve this, we propose a Target-Aware Prompting Strategy, as illustrated in Fig. 4(a). We construct a following composite prompt P by embedding the expression E into a query template:

$$P = \text{“Question: Describe the scene context around the } [E]\text{? Answer:”} \quad (1)$$

During inference, the VLM takes the image I and prompt P to generate the caption $C = \text{VLM}_{\text{frozen}}(I, P)$. This strategy forces the VLM to act as a semantic densifier, expanding the local expression with rich, target-relevant details (e.g., attributes and background) rather than generic descriptions.

Prior Encoding. Finally, to align this explicit prior with the RIS feature space, we encode the generated caption C using the **shared text encoder** (from Section 3.2). This yields the global semantic features $F_c \in \mathbb{R}^{L_c \times D}$. Sharing the encoder ensures that the global prior F_c and local expression F_l reside in a unified semantic space, facilitating effective interaction in the subsequent module.

3.4 Dual-Path Global Context Injection

We propose the Dual-Path Context Injection Module to effectively integrate the Global Semantic Prior, simultaneously infusing the structured prior F_c into both the local visual (F_v) and linguistic (F_l) streams.

Global-to-Visual Injection. To enrich dense visual features F_v with high-level semantics, we aggregate global context from the caption prior F_c via cross-attention, using F_v as the Query and F_c as Key and Value:

$$\hat{F}_v = \text{Softmax}\left(\frac{F_v F_c^\top}{\sqrt{D}}\right) F_c \tag{2}$$

where D is the scaling factor. Through this operation, each local visual region absorbs the most relevant global semantics (e.g., object relationships) defined in the caption, yielding the enriched representation $\hat{F}_v$.

Global-to-Linguistic Injection. Simultaneously, to densify the local linguistic embeddings F_l, we inject the global background information. We perform a symmetric interaction by treating the linguistic features F_l as the Query:

$$\hat{F}_l = \text{Softmax}\left(\frac{F_l F_c^\top}{\sqrt{D}}\right) F_c \tag{3}$$

This produces a context-aware linguistic representation $\hat{F}_l$, effectively expanding the semantic receptive field of the original expression.

Adaptive Semantic Gating. To robustly integrate the global prior while filtering potential noise from the generated caption, we introduce a learnable **Adaptive Gating Mechanism**. We compute a gate coefficient vector $\mathbf{g}_v$ to dynamically modulate the injection intensity:

$$g_v = \sigma(W_v[F_v; \hat{F}_v] + b_v), \quad g_l = \sigma(W_l[F_l; \hat{F}_l] + b_l) \tag{4}$$

where σ denotes the Sigmoid function, and $[\cdot;\cdot]$ represents concatenation. Finally, we apply these gates to obtain the context-enriched representations:

$$F_v^r = F_v + g_v \odot \hat{F}_v, \quad F_l^r = F_l + g_l \odot \hat{F}_l \tag{5}$$

where $\odot$ represents element-wise multiplication. Crucially, these refined features F_v^r and F_l^r effectively integrate fine-grained local details with holistic scene semantics. Acting as semantically enriched representations, they replace the original features to robustly guide the subsequent decoding process.

3.5 Mask Prediction and Training Objectives

Mask Prediction. Following the dual-path injection, the context-enriched representations F_v^r and F_l^r are fed into the standard segmentation decoder of the base framework (e.g., LAVT [33]). The decoder performs the final pixel-level classification to predict the binary segmentation mask $M \in \{0,1\}^{H \times W}$.

Semantic Consistency Constraint. To enforce the visual backbone's comprehension of the Global Semantic Prior, we introduce a Semantic Consistency Constraint that aligns the high-level visual summary with the caption.

Formally, we define the *visual summary* vector $\mathbf{v}_{global}$ by performing global average pooling on the refined visual features:

$$\mathbf{v}_{global} = \frac{1}{N_v} \sum_{i=1}^{N_v} F_v^r[i] \tag{6}$$

Correspondingly, let $\mathbf{c}_{global}$ denote the global [CLS] token of the caption features F_c, representing the semantic summary. We maximize their alignment in the shared semantic space by minimizing the Cosine distance:

$$\mathcal{L}_{cons} = 1 - \frac{\mathbf{v}_{global} \cdot \mathbf{c}_{global}}{\|\mathbf{v}_{global}\| \cdot \|\mathbf{c}_{global}\|} \tag{7}$$

This regularization term penalizes the model when the visual features deviate from the global semantic prior, effectively preventing semantic tunnel vision.

Total Objective. The entire framework is optimized end-to-end. The total objective is a weighted combination of the standard segmentation loss $\mathcal{L}_{seg}$ (Binary Cross-Entropy + Dice) and our global consistency loss:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda \mathcal{L}_{cons} \tag{8}$$

where λ controls the relative contribution of the global semantic consistency constraint, balancing pixel-level supervision and high-level semantic alignment.

4 Experiments

4.1 Experimental Settings

Datasets and Metrics. We evaluate our method on three standard benchmarks: **RefCOCO** [15], **RefCOCO+** [15], and **RefCOCOg** [35]. RefCOCO+ is particularly noted for prohibiting absolute spatial terms, forcing reliance on appearance attributes. RefCOCOg contains significantly longer and more complex expressions, serving as a rigorous test for global semantic understanding. Following standard protocols [33], we use the UNC partition for RefCOCO/RefCOCO+

Table 1. Comparison with state-of-the-art methods on RefCOCO, RefCOCO+, and RefCOCOg. Results are reported in mIoU (%). **Ours** denotes CGA-Net integrated into three baselines. Notably, the significant gains on RefCOCOg demonstrate the effectiveness of our Global Semantic Prior in complex scenarios.

Method	Backbone	RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val-u	test-u
I. State-of-the-Art Methods									
CRIS [30]	Res-101	70.47	73.18	66.10	62.27	68.08	53.68	59.87	60.36
VLT [12]	Swin-B	72.96	75.96	69.60	63.53	68.43	56.92	63.49	66.22
SeqTR [37]	Swin-B	71.70	73.31	69.82	63.04	66.73	58.97	64.69	65.74
VG-LAW [29]	ViT-B	75.05	77.36	71.69	66.61	70.30	58.14	65.36	65.13
LISA-7B [17]	SAM-H	74.90	79.10	72.30	65.10	70.80	58.10	67.90	70.60
II. Validation on Pixel-level Paradigm									
LAVT [33] (Baseline)	Swin-B	74.46	76.89	70.94	65.81	70.97	59.23	63.34	63.62
LAVT + Ours	Swin-B	**76.82**	**79.15**	**72.88**	**68.14**	**73.22**	**61.18**	**65.85**	**66.05**
III. Validation on Sequence-Generation Paradigm									
PolyFormer-B [20]	Swin-B	76.94	78.49	74.83	72.15	75.71	66.73	71.15	71.17
PolyF. + Ours	Swin-B	**79.12**	**80.55**	**76.78**	**74.02**	**77.58**	**68.35**	**73.28**	**73.45**
IV. Validation on Implicit Context Paradigm									
VATEX [25]	Swin-B	78.16	79.64	75.64	70.02	74.41	62.52	69.73	70.58
VATEX + Ours	Swin-B	**80.58**	**81.86**	**77.42**	**72.65**	**76.92**	**64.58**	**72.83**	**73.53**

and the UMD partition for RefCOCOg. We adopt the **Mean Intersection-over-Union (mIoU)** as the primary evaluation metric.

Implementation Details. Crucially, to ensure training efficiency, we employ an **offline caching strategy**: the global semantic priors (F_c) are pre-computed and cached before training. This strategy avoids repetitive VLM inference, ensuring that incorporating the global prior imposes negligible computational overhead during the training phase.

4.2 Results

Quantitative Results. Table 1 reports the comparison with state-of-the-art methods. CGA-Net establishes new state-of-the-art results across all datasets. **(1) Universality:** Whether integrated into pixel-based (LAVT), sequence-based (PolyFormer), or context-aware (VATEX) architectures, our module consistently yields performance gains. **(2) Complex Scenarios:** On the challenging **RefCOCOg (UMD-test)**, CGA-Net outperforms the strong PolyFormer baseline by **+2.28%**. This significant boost confirms that our Global Semantic Prior effectively resolves the semantic bottleneck in complex scenes.

Qualitative Analysis. Fig. 5 presents visual comparisons. The baseline (LAVT) often fails on relational queries, resulting in incorrect segmentation, whereas CGA-Net leverages generated captions to disambiguate targets, demonstrating the effectiveness of structured semantic guidance.

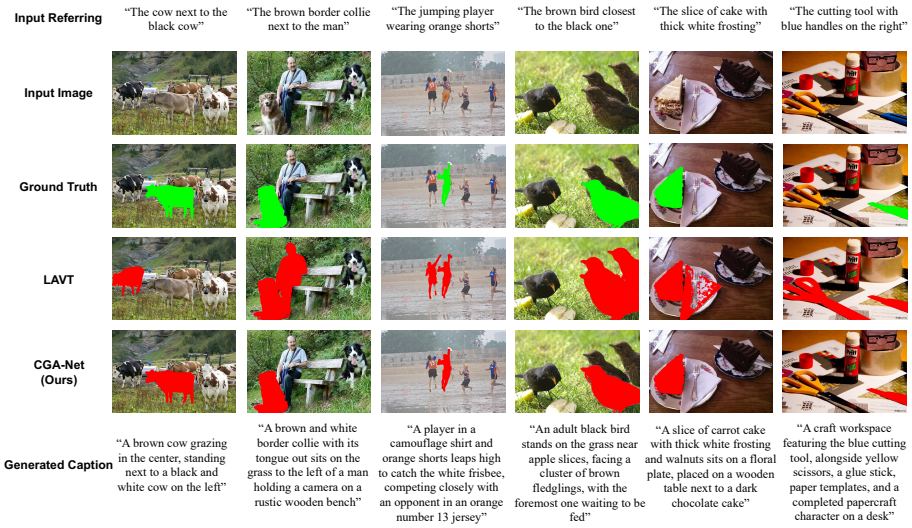

Fig. 5. Qualitative results on RefCOCO(+/g) datasets. The generated captions provide structured cues that help CGA-Net disambiguate targets in complex scenes.

Table 2. Precision analysis (Pr@X) on RefCOCO val. CGA-Net shows superior boundary quality at high IoU thresholds (Pr@0.9), significantly outperforming baselines by effectively leveraging the Global Semantic Prior.

Method	Pr@0.5	Pr@0.7	Pr@0.9	mIoU
LAVT [33]	84.46	75.28	34.30	74.46
ReLA [19]	85.92	77.71	34.99	75.61
VATEX [25]	88.12	82.54	45.11	78.17
CGA-Net (Ours)	**89.45**	**84.10**	**47.20**	**80.58**

4.3 Ablation Studies

We conduct component-wise ablation studies on the RefCOCO val set using LAVT-SwinB as the baseline. To ensure fair comparison, all ablation variants follow the exact same training schedule as the main experiments.

Effectiveness of Core Components. Table 3 quantitatively validates the contribution of our proposed modules. The baseline LAVT achieves 74.46% mIoU.

1. Impact of Caption Integration.

The **Dual-Path Context Injection Module** (Row 2) yields the largest gain. By utilizing captions as a **Global Semantic Prior**, it **supplements** sparse expressions with **structured scene descriptions**, bridging the **informational gap** to distinguish visually similar targets.

2. Benefit of Adaptive Gating. Adding **Adaptive Gating** (Row 3) further improves performance by acting as a semantic filter. It dynamically mod-

Table 3. Component-wise ablation. Analyzing the impact of Dual-Path Injection, Gating, and Consistency Loss.

Row	Dual-Path	Gate	Consistency	mIoU
1	-	-	-	74.46
2	✓	-	-	76.10
3	✓	✓	-	76.45
4	✓	✓	✓	**76.82**

Table 4. Impact of Semantic Source. Structured captions outperform implicit features.

Source	Type	mIoU
None	-	74.46
Global Vis. Feat.	Implicit	75.12
Obj. Tags	Unstruct.	75.95
Caption (Ours)	**Structured**	**76.82**

Table 5. Comparison of Prompting Strategies. Target-aware prompting reduces noise and maximizes gains.

Strategy	Template	mIoU
Generic	"Describe this image."	75.60
Simple Ref	"A photo of [Ref]"	76.15
Target-Aware	"Context around [Ref]"	**76.82**

ulates the prior's influence, mitigating the impact of caption hallucinations and ensuring injected context does not mislead local visual alignment.

3. Role of Consistency Loss. Finally, enforcing the **Semantic Consistency Constraint** (Row 4) achieves the best results. By aligning the visual global summary with the caption, it forces the encoder to learn language-aware representations, ensuring global coherence while focusing on local details.

Impact of Semantic Source. As shown in Table 4, we investigate different forms of semantic priors. Implicit visual features (e.g., CLIP image embeddings) provide limited gains (+0.66%) as they remain entangled in the visual feature space. Unstructured object tags (e.g., a list of detected objects) improve performance to 75.95% but fail to capture relational context. In contrast, our Structured Caption achieves the highest mIoU (76.82%), verifying that the linguistic structure (relationships and attributes) is key to bridging the informational gap.

Impact of Prompting Strategy. Table 5 compares different prompting templates used for caption generation. Generic prompts (e.g., "Describe this image") often lead to hallucinations or irrelevant descriptions that distract the model. Conversely, our **Target-Aware Prompt** ("Context around [Ref]") effectively directs the VLM's attention to the specific region of interest. This acts as a focused "semantic densifier," resulting in the best performance.

5 Conclusion

In this paper, we propose CGA-Net to resolve the inherent insufficiency of global context in Referring Image Segmentation. By leveraging a Global Semantic Prior,

our Dual-Path Context Injection Module bridges the informational gap between sparse queries and dense visual scenes. Extensive evaluations confirm that CGA-Net establishes a new state-of-the-art across RefCOCO benchmarks, effectively resolving referential ambiguities where traditional internal-alignment methods fail. Future work will explore end-to-end fine-tuning of the VLM component to further enhance prior adaptability and generalization capability.

Acknowledgements This work was supported by JSPS KAKENHI Grant Number 23K28164, 24H00733 and JST CREST Grant Number JPMJCR22D1.

1 2 3 4

References

1. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
3. Chen, J., Deguchi, D., Li, D., Zheng, X., Ito, S., Murase, H., Fan, Q.: Semantic-centric alignment for zero-shot panoptic segmentation with limited data. International Journal of Computer Vision **134**(3), 122 (2026)
4. Chen, J., Deguchi, D., Zhang, C., Zheng, X., Murase, H.: Frozen is better than learning: A new design of prototype-based classifier for semantic segmentation. Pattern Recognition **152**, 110431 (2024)
5. Chen, J., Li, D., Yi, C., Zheng, X., Seigo, I., Murase, H., Deguchi, D.: Semantic-driven distillation for semantic segmentation with unknown classes. In: 2025 IEEE 28th International Conference on Intelligent Transportation Systems (ITSC). pp. 2655–2662. IEEE (2025)
6. Chen, J., Quan, Z., Zhang, C., Zheng, X., Deguchi, D., Murase, H.: Clip-to-seg distillation for zero-shot semantic segmentation. IEEE Transactions on Circuits and Systems for Video Technology (2025)
7. Chen, J., Zheng, X., Li, D., Yi, C., Ito, S., Paudel, D.P., Gool, L.V., Murase, H., Deguchi, D.: Split matching for inductive zero-shot semantic segmentation. In: 36th British Machine Vision Conference 2025, BMVC 2025, Sheffield, UK, November 24-27, 2025. BMVA (2025)
8. Chen, J., Zheng, X., Paudel, D.P., Van Gool, L., Murase, H., Deguchi, D.: Bix-former: A robust framework for maximizing modality effectiveness in multi-modal semantic segmentation. arXiv preprint arXiv:2506.03675 (2025)
9. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)

10. Dai, M., Cheng, W., Liu, J.j., Yang, S., Cai, W., Sun, Y., Yang, W.: Deris: Decoupling perception and cognition for enhanced referring image segmentation through loopback synergy. arXiv preprint arXiv:2507.01738 (2025)
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
12. Ding, H., Liu, C., Wang, S., Jiang, X.: Vlt: Vision-language transformer and query generation for referring segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(6), 7900–7916 (2022)
13. Huang, S., Hui, T., Liu, S., Li, G., Wei, Y., Han, J., Liu, L., Li, B.: Referring image segmentation via cross-modal progressive comprehension. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10488–10497 (2020)
14. Hui, T., Liu, S., Huang, S., Li, G., Yu, S., Zhang, F., Han, J.: Linguistic structure guided context modeling for referring image segmentation. In: European conference on computer vision. pp. 59–75. Springer (2020)
15. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 787–798 (2014)
16. Kim, S., Kang, M., Park, J.: Risclip: Referring image segmentation framework using clip. CoRR (2023)
17. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9579–9589 (2024)
18. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
19. Liu, C., Ding, H., Jiang, X.: Gres: Generalized referring expression segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23592–23601 (2023)
20. Liu, J., Ding, H., Cai, Z., Zhang, Y., Satzoda, R.K., Mahadevan, V., Manmatha, R.: Polyformer: Referring image segmentation as sequential polygon generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18653–18663 (2023)
21. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
22. Liu, S., Hui, T., Huang, S., Wei, Y., Li, B., Li, G.: Cross-modal progressive comprehension for referring segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(9), 4761–4775 (2021)
23. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
24. Luo, G., Zhou, Y., Sun, X., Cao, L., Wu, C., Deng, C., Ji, R.: Multi-task collaborative network for joint referring expression comprehension and segmentation. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 10034–10043 (2020)

25. Nguyen-Truong, H., Nguyen, E.R., Vu, T.A., Tran, M.T., Hua, B.S., Yeung, S.K.: Vision-aware text features in referring image segmentation: From object understanding to context understanding. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4988–4998. IEEE (2025)
26. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18082–18091 (2022)
27. Ren, Z., Huang, Z., Wei, Y., Zhao, Y., Fu, D., Feng, J., Jin, X.: Pixellm: Pixel reasoning with large multimodal model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26374–26383 (2024)
28. Shang, C., Song, Z., Qiu, H., Wang, L., Meng, F., Li, H.: Prompt-driven referring image segmentation with instance contrasting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4124–4134 (2024)
29. Su, W., Miao, P., Dou, H., Wang, G., Qiao, L., Li, Z., Li, X.: Language adaptive weight generation for multi-task visual grounding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10857–10866 (2023)
30. Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T.: Cris: Clip-driven referring image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11686–11695 (2022)
31. Wu, J., Jiang, Y., Sun, P., Yuan, Z., Luo, P.: Language as queries for referring video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4974–4984 (2022)
32. Yan, B., Jiang, Y., Wu, J., Wang, D., Luo, P., Yuan, Z., Lu, H.: Universal instance perception as object discovery and retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15325–15336 (2023)
33. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: Lavt: Language-aware vision transformer for referring image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18155–18165 (2022)
34. Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., Berg, T.L.: Mattnet: Modular attention network for referring expression comprehension. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1307–1315 (2018)
35. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: European Conference on Computer Vision (ECCV). pp. 69–85. Springer (2016)
36. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)
37. Zhu, C., Zhou, Y., Shen, Y., Luo, G., Pan, X., Lin, M., Chen, C., Cao, L., Sun, X., Ji, R.: Seqtr: A simple yet universal network for visual grounding. In: European Conference on Computer Vision. pp. 598–615. Springer (2022)
38. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. NeurIPS (2023)