Phrase-Instance Alignment for Generalized Referring Segmentation

Supplementary Material

1. Details about Datasets and Metrics

1.1. gRefCOCO

Proposed by Liu *et al*. [9], the gRefCOCO dataset consists of 19,994 images of 60,287 distinct instances described by 278,232 language expressions. These annotations include 80,022 multi-target and 32,202 no-target samples.
For evaluation on gRefCOCO dataset, we use the following metrics:

- **cIoU** (cumulative Intersection over Union): Measures the total pixel intersection over the total pixel union across the validation split, offering a holistic view of segmentation performance.
- **gIoU** (generalized Intersection over Union): Averages the per-image IoU across all samples to evaluate segmentation precision at the image level. For no-target samples, the per-image IoU is 1 if we predict correctly a no-target sample and 0 in the other cases.
- **N-acc** (no-target accuracy): Quantifies the accuracy of no-target identification, i.e., how well the model identifies cases where no target is present.

1.2. Ref-ZOM

Proposed by Hu*et al*. [5], the Ref-ZOM is similar to gRefCOCO, introducing the one-to-one, one-to-many, and one-to-zero referring expressions. These cases correspond to the single-target, multi-target, and no-target samples in gRefCOCO, respectively. Ref-ZOM dataset consists of 55,078 images of 74,942 annotated instances, which includes 41,842 annotated objects under one-to-many settings, 11,937 one-to-zero samples, and 42,421 one-to-one objects.

For Ref-ZOM, gIoU and cIoU are substituted to the equivalent metrics in RES, mIoU and oIoU. However, different from gRefCOCO, mIoU and gIoU only count for one-to-one and one-to-many samples. For one-to-zero samples, we use the accuracy (acc) metric, which measures the classification performance on empty-target expressions.

1.3. Instances ground truth

While the metrics to measure the performance for GRES task are based on the binary ground-truth mask (that combine all instances together), both gRefCOCO [9] and Ref-ZOM [5] datasets provide instance mask annotations for each expression. We utilize these annotations for our instance supervision. This supervision is not external data but comes directly from the gRefCOCO and Ref-ZOM datasets, where instance-level annotations are already provided and accessible to all methods. What distinguishes our approach is that we explicitly model and supervise individual object instances, which previous methods have not done, largely because they lack the instance-aware design needed to make use of this information. Our work aims to fill that gap. We also note that other recent methods have started leveraging richer signals as well—for example, CoHD [15] uses instance-level information for object counting, and LLM-based models [8, 23] often rely on large-scale external datasets. While our supervision setup is different from earlier works, we believe it's a principled use of existing data more effectively.

2. Implementation Details

Our model is optimized using AdamW [13] optimizer with the initial learning rate of 10^{-4} and linear decreasing to 10^{-6} after 20 epochs. Following Mask2Former, , the coefficients of instance supervision are set as $\lambda_{score} = 2, \lambda_{mask} = 5.\lambda_{phrase} = 1$. Following ReLA [9], we set $\lambda_{merged} = 1$ and $\lambda_{nt} = 0.1$. We set $\lambda_{phrase} = 1$ and $\lambda_{inst} = 1$ as default hyperparameters. BERT-base-uncased [12] is used as the Text Encoder to extract language features and the Pixel Decoder comprises 6 layers of Deformable Transformer, following Mask2Former [2].

3. Analysis

3.1. Phrase-Object Alignment Loss

The value of $\mathcal{L}_{\text{phrase}}$ directly measures the similarity between object queries and their corresponding phrase embeddings in terms of cosine distance. To better understand how closely these features align, we compute the converged values of $\mathcal{L}_{\text{phrase}}$ on the gRefCOCO validation set under two selection strategies:

- **Oracle setting (ground-truth known):** We assume access to ground-truth instance masks and use Hungarian matching to select the top-M predicted object queries—one for each ground-truth instance. This mirrors the matching logic applied during training.
- **Natural setting (inference-time behavior):** No ground-truth information is accessed. Instead, we select predicted object queries based on their relevance scores, retaining those with $\hat{p}_i > 0.5$. This simulates the model's natural selection process at inference time.

$\mathcal{L}_{\text{phrase}}$	**Average**	**Median**	**Min**	**Max**
Oracle setting	0.194	0.135	0.02	0.852
Natural setting	0.142	0.105	0.03	0.697

These values show that while the alignment is meaningful, it is far from perfect—confirming that object queries and phrase embeddings do not collapse into identical representations. The non-negligible distances reflect healthy variation across queries and help preserve distinct semantics, which is essential for accurate instance discrimination.

3.2. Ablation on loss coefficients

Since we adopt the framework from Mask2Former [2] and the concept of no-target predictor from [9], we keep their default hyperparameters. We report the ablation study about the instance supervision λ_{inst} and λ_{phrase} in the table below. As we can observe, the balance coefficient between them yields the best performance onthe gRefCOCO validation set.

λ_{inst}	λ_{phrase}	cIoU	N-acc.
1	1	**68.94**	**79.72**
1	2	67.63	73.38
1	0.5	68.45	77.23
2	1	68.12	77.92
0.5	1	67.65	76.26

Table 1. Ablation study on loss coefficients.

3.3. Size and performance

We provide detailed comparisons of performance, model size and FLOPs on table 2. Our model, *InstAlign*, achieves the best accuracy (gIoU: 74.51, cIoU: 68.94, N-acc.: 79.72) while maintaining a balanced computational cost (230M parameters, 0.235T FLOPs).

Method	Backbone	Parameters	T-FLOPs	gIoU	cIoU	N-acc.
DMMI	Swin-B	341M	0.392T	62.68	62.77	53.20
ReLA	Swin-B	226M	0.131T	63.60	62.42	56.37
CoHD	Swin-B	248M	0.133T	68.42	65.17	63.68
InstAlign	Swin-B	230M	0.235T	74.51	68.94	79.72

Table 2. Performance and efficiency comparison with previous SOTA method on gRefCOCO validation set.

Our primary reason for using the Swin-B/BERT-base configuration was to ensure a direct and fair comparison with the most relevant GRES methods, including ReLA, LQMFormer, and CoHD, which all adopt the same Swin-B and BERT-base backbones. Using this configuration enables a clear evaluation of the methodological contributions introduced by our approach. To demonstrate generalization across architectures, we also experiment with additional visual and text encoders, including Swin-L and RoBERTa-base. Results are summarized in Table 3.

Visual Encoder	Text Encoder	cIoU	gIoU	N-acc
Swin-B (default)	BERT	68.94	74.34	79.72
Swin-B	RoBERTa	68.72	74.51	79.63
Swin-L	BERT	**69.04**	**74.82**	**80.25**

Table 3. Performance comparison across different backbone configurations.

3.4. Design Rationale for Instance Aggregation

Our instance aggregation module builds on the instance-level object queries and their relevance scores, providing a lightweight refinement that determines how these predictions are merged into the final mask and directly influences segmentation accuracy. As shown in Table 3(b) main, directly selecting high-confidence instances for the final mask achieves 66.67% cIoU and 72.96% N-acc, demonstrating that our instance-aware approach enables the model to predict individual instances independently with reasonable accuracy. However, by incorporating our Instance Aggregation (IA), we achieve an additional +2.0% cIoU and +6.0% N-acc, showing that IA's soft-assignment approach mitigates the risk of missing relevant instances or including irrelevant ones, leading to more precise segmentation.

One key design choice in IA is performing aggregation on the logits (pre-sigmoid values) of instances rather than their probability scores. This is crucial because logits preserve a linear combination space, allowing the model to better maintain the relative importance of each instance. In contrast, probability values are bounded between 0 and 1, which compresses differences between instances, potentially skewing the merging process. By merging logits instead of probabilities, we avoid this distortion and retain finer segmentation details.

One of the key motivations for choosing PReLU activation is to weaken the influence of negative logits, which typically correspond to irrelevant or background regions. By applying PReLU, we allow the model to suppress negative logits more effectively, enhancing the overall merging process by reducing the contribution of less relevant or distracting instances. While this approach improves the results empirically, understanding why weakening negative logits benefits merging requires further investigation, possibly involving an analysis of how instance logits interact during the combination process.

3.5. Number of Object Queries

In both gRefCOCO and Ref-ZOM, a single referring expression can correspond to many distinct object instances—up to 18 in the most extreme cases. Because our training uses Hungarian matching for one-to-one supervision, the number of object queries N must be sufficiently large to accommodate all potentially referred instances. Otherwise, some ground-truth objects would remain unmatched.

We also examined settings with a reduced query capacity. Using only $N = 10$ queries leads to a clear performance drop (cIoU 64.15%), illustrating that insufficient query slots limit the model's ability to represent all relevant instances. In our final model, we adopt $N = 100$, which provides a safe margin and consistently stable performance.

Dataset	gRefCOCO				Ref-ZOM	
	train	val	testA	testB	train	test
M	18	14	16	14	18	13

Table 4. Maximum number of referred object instances (M) per expression across datasets.

4. Grounding DINO + SAM

To contextualize the difficulty of Generalized Referring Expression Segmentation (GRES), we evaluate a strong detection-segmentation pipeline using Grounding DINO-base for text-conditioned object detection followed by SAM2.1-Hiera-Large for mask extraction. As shown in Table 5, this pipeline performs reasonably well on single-object RES (RefCOCO-val), but its performance drops significantly on multi-object and no-target cases in gRefCOCO. This highlights the inherent challenge of GRES and the importance of instance-aware reasoning in our approach.

gRefCOCO-val			RefCOCO-val
cIoU	gIoU	N-acc	oIoU
33.2	42.0	34.9	67.0

Table 5. Performance of the Grounding DINO-base + SAM2.1-Hiera-Large baseline.

Method	Backbone	RefCOCO			RefCOCO+			RefCOCOg	
		val	test A	test B	val	test A	test B	val	test
VLT [4]	Darknet-53	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65
ReSTR [7]	ViT-B/16	67.22	69.30	64.45	55.78	60.44	48.27	-	-
CRIS [22]	ResNet-101	70.47	73.18	66.10	62.27	68.08	53.68	59.87	60.36
LAVT [24]	Swin-B	72.73	75.82	68.79	62.14	68.38	55.10	61.24	62.09
ReLA [9]	Swin-B	73.82	76.48	70.18	66.04	71.02	57.65	65.00	65.97
DMMI [5]	Swin-B	74.14	77.13	70.16	63.98	69.73	57.03	61.98	63.46
LQMFormer [18]	Swin-B	74.16	76.82	71.04	65.91	71.84	57.59	64.73	66.04
CGFormer [21]	Swin-B	74.75	77.30	70.64	64.54	71.00	57.14	62.51	64.68
PolyFormer† [10]	Swin-B	74.82	76.64	71.06	67.64	72.89	59.33	67.76	69.05
LISA-7B† [8]	SAM-ViT-H	74.1	76.5	71.1	62.4	67.4	56.5	66.4	68.5
GSVA-7B† [23]	SAM-ViT-H	77.2	78.9	73.5	65.9	69.6	59.8	72.7	73.3
MagNet‡ [3]	Swin-B	76.55	78.27	72.15	68.10	73.64	61.81	67.79	69.29
CGFormer [21]	Swin-B	76.93	78.70	73.32	68.56	73.76	61.72	67.57	67.83
Prompt-RIS† [19]	SAM-ViT-B/16	76.36	80.37	72.29	67.06	73.58	58.96	64.79	67.16
VATEX‡ [17]	Swin-B	**81.53**	**82.75**	**79.66**	**74.61**	**78.75**	**68.52**	75.54	76.40
InstAlign(**Ours**)‡	Swin-B	78.05	80.02	75.03	68.32	72.85	60.15	73.47	75.73

Table 6. Comparison with SOTA methods in RES task using the oIoU metric. ‡ indicates combining the train splits from these 3 datasets with test images removed to prevent data leakage. † indicates using additional data beyond RefCOCO, RefCOCO+, and G-Ref.

5. Instance-aware Supervision details

Our key design focuses on integrating input language expression into a query-based instance segmentation framework and guiding it to identify and segment only instances described in the input text. Here, we adopt Mask2Former as our instance segmentation framework due to its effectiveness and efficiency.

5.1. Feature Extraction

To extract the text information, we adopted BERT-based model [12] to embed the input expression into high-level word features $T_0 \in \mathbb{R}^{L\times C}$, where C and L denotes the number of channels and the length of the expression, respectively. We use an encoder-decoder architecture to extract the multi-scale pixel features $\mathcal{V} = \{V_i \in \mathbb{R}^{C\times H_i\times W_i}\}_{i=1}^{4}$ from the input image. Here, H_i and W_i denote the height and the width of the feature maps in the i-th stage, respectively. As in [2], these visual features V_i are obtained from the Pixel Decoder [2, 25].

5.2. Transformer Decoder

Our transformer decoder is adopted from Mask2Former [2] and ReLA [9]. To process an input query, the model operates on a set of N end-to-end trained object queries, alongside extracted text and visual features. Through a sequence of K transformer blocks, illustrated in Fig. 1, it progressively refines these queries by integrating visual and textual features while simultaneously enriching text representations with object-aware information.

More specifically, the k-th transformer block takes the pixel features $V_k \in \mathbb{R}^{H_k\times W_k\times C}$, text features $T_{k-1} \in \mathbb{R}^{L\times C}$, and object query $Q_{k-1} \in \mathbb{R}^{N\times C}$ as input and outputs refined object query Q_k and text features T_k. Visual features on specific scales are input to different blocks in a round-robin manner, and where V_k denotes the visual features input to the

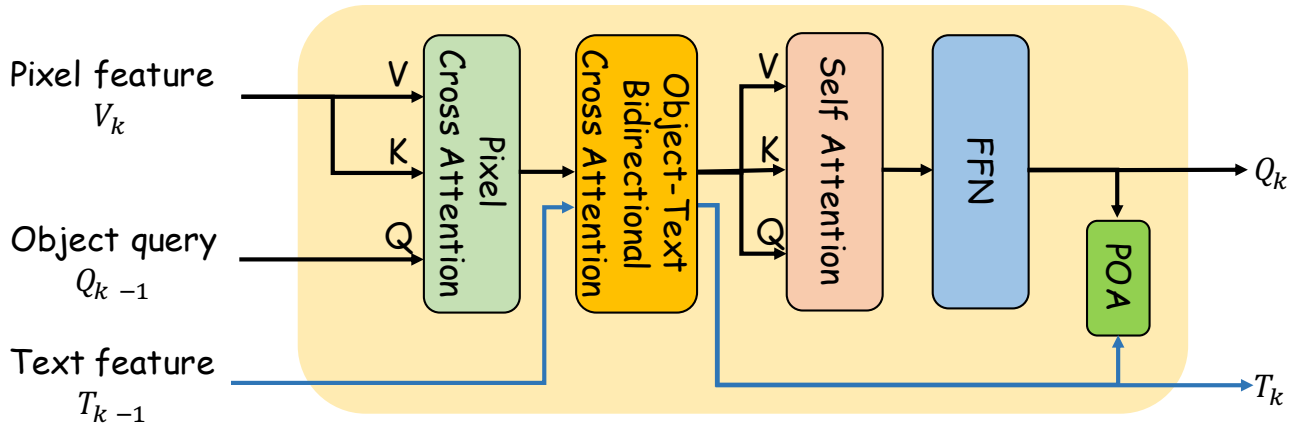

Figure 1. **Transformer Decoder**. Object query and Text feature are refined by each other and Pixel feature.

k-th transformer block. The bidirectional object-text cross-attention layer allows both the text features and the object queries to be transformed on the basis of information from both sides. To obtain refined object queries Q_k, we sequentially pass the object queries Q_{k-1} through a cross-attention layer with visual features V_k, an object-text cross-attention layer with text features T_{k-1}, a self-attention layer and an FFN layer. Simultaneously, the text features T_{k-1} are passed through the same object-text cross attention layer to produce the refined text features T_k. We note that this refining text feature is a minor enhancement designed to enrich text representations, and we do not claim contributions from this.

5.3. Prediction Heads

Given a refined object query $Q_K \in \mathbb{R}^{N\times C}$, we first predict the probability $\hat{p} \in \mathbb{R}^N$ of these instances being related to the expression input via a simple MLP layer. Similar, N instance masks $\hat{s} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times N}$ associated to this object query can be computed by:

$$\hat{s} = V_4 \cdot Q_K^\top, \tag{1}$$

where $V_4 \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C}$ is the pixel features at the highest-resolution scale extracted from the Pixel Decoder (Sec. 5.1)

5.4. Instance Supervision

The ground truth segmentation $\mathcal{M}_{gt} \in \mathbb{R}^{H\times W}$ is constructed by combining a set of M ground-truth instance segments $s = \{s_i\}_{i=1}^M$. To supervise our predictions effectively, we adopt the instance supervision from Mask2Former [2] to establish a one-to-one correspondence between the predicted instances and the ground-truth instances. Specifically, we find the optimal set of prediction indices $\omega^* = \{\omega_i\}_{i=1}^M$ that minimizes the following matching cost:

$$\omega^* = \arg\min_{\omega\subseteq[N]} \sum_{i=1}^{M} \mathcal{L}_{\text{match}}(\omega_i, i), \tag{2}$$

where the per-instance matching cost $\mathcal{L}_{\text{match}}$ for i-th predicted instance and j-th ground-truth instance is defined as:

$$\mathcal{L}_{match}(i,j) = \lambda_{score}\mathcal{L}_{score}(\hat{p}_i, 1) + \lambda_{mask}\mathcal{L}_{mask}(\hat{s}_i, s_j). \tag{3}$$

Here, $\mathcal{L}_{score}$ is the BCE loss that promotes high confidence for relevant instances, $\mathcal{L}_{mask}$ is the combination of dice loss [20] and BCE loss to improve mask consistency. The hyper-parameters $\lambda_{score}, \lambda_{mask}$ control the balance of each component in the matching cost.

After determining the optimal match, the instance loss $\mathcal{L}_{inst}$ is used to supervise the model for both matched and unmatched samples. The total instance loss is given by:

$$\mathcal{L}_{inst} = \sum_{i=1}^{M} \mathcal{L}_{match}(\omega_i^*, i) + \sum_{i=1, i\notin\omega^*}^{N} \mathcal{L}_{score}(\hat{p}_i, 0), \tag{4}$$

where the first term ensures that for each matched instance, the model minimizes the combined matching loss while the second term penalizes unmatched instances by encouraging their relevance scores $\hat{p}_i$ to approach zero, indicating irrelevance to the expression. It is important to note that, similar to prior works [1, 2], the number of object queries N does not need to equal the actual number of relevant objects M.

6. More quantitative results

6.1. Ref-ZOM

Table 1 showcases the results of our method compared to state-of-the-art methods across one-to-one, one-to-many, and one-to-zero cases in the Ref-ZOM dataset. Our model demonstrates significant improvements across all scenarios. In one-to-one cases, *InstAlign* surpasses prior SOTA with a margin of 0.93% and 2.58% in terms of oIoU and mIoU, respectively, showcasing our precision in identifying individual objects. In one-to-many cases, our model outperforms DMMI [5] with a large margin of 4.47% and 4.65% in oIoU and mIoU. For the no-target scenario, the Acc score of 94.23% indicates our model's ability to accurately determine when no valid target is present compared to previous approaches.

These results highlight the effectiveness of explicitly predicting relevant instances and instance aggregation, enabling accurate segmentation for complex expressions.

Method	One-to-One		One-to-Many		Overall Targets		One-to-Zero
	oIoU	mIoU	oIoU	mIoU	oIoU	mIoU	Acc
MCN [14]	52.09	53.14	58.04	57.21	55.05	54.70	75.81
CMPC [11]	52.46	52.89	60.23	60.27	56.19	55.72	77.01
VLT [4]	59.07	58.96	61.42	62.79	60.21	60.43	79.26
LAVT [24]	63.21	64.56	65.69	65.14	64.45	64.78	83.11
DMMI [5]	65.43	66.83	72.20	70.44	68.77	68.21	87.02
InstAlign (**Ours**)	**66.36**	**68.41**	**76.77**	**75.09**	**71.13**	**70.81**	**94.23**

Table 7. Quantitative comparison across 3 cases of Ref-ZOM dataset.

6.2. Performance on Traditional RES

While our primary focus is addressing multi-target scenarios in Generalized Referring Expression Segmentation, we also evaluate our model on the traditional RES task to provide a broader performance perspective. We follow MagNet [3] to combine the train splits from 3 datasets RefCOCO, RefCOCO+ [6], and RefCOCOg [16] with test images removed to prevent data leakage. As shown in Table 2, while *InstAlign* is designed for GRES, it also performs competitively on traditional RES benchmarks.

This indicates that our instance-aware reasoning and phrase-object alignment are beneficial even in traditional RES, suggesting broader applicability.

7. More Qualitative Results

7.1. Ref-ZOM

In this subsection, we present qualitative results on the Ref-ZOM dataset, which is designed to test the ability of models to handle complex referring expressions that describe multiple objects or involve intricate spatial relationships. As shown in Fig. 2, our model effectively segments the objects based on the input expressions, including multi-object scenarios and fine-grained localization.

7.2. Failure Case Analysis

While our approach achieves promising results in many scenarios, there are certain cases where the model struggles. As shown in Fig. 3, these failure cases primarily arise in three main scenarios: (1) Ambiguous expressions, (2) Fine-grained object distinction failures, and (3) Complex spatial reasoning issues. For instance, expressions like "the second elephant from us" lead to incorrect segmentations because the model has difficulty counting the target from other similar objects in the scene. Similarly, expressions that involve hierarchical spatial relationships, such as "the man in the suit and blue tie behind the old woman's arm,".

A particularly difficult case occurs when segmenting "51" This requires the model to identify a small, specific region (the number on the shirt) and associate it with the correct individual. Since existing segmentation networks primarily rely on global object shape and texture, fine-grained visual cues like text remain challenging. Future improvements could involve integrating OCR-based feature extraction or multi-level attention mechanisms to better capture textual details within object regions.

Figure 2. **Qualitative results on Ref-ZOM dataset.** Our model successfully segments objects based on challenging referring expressions, handling complex multi-object scenarios. Best viewed in color.

Despite these challenges, our model shows strong robustness in the majority of cases, and further research into handling ambiguities and improving spatial reasoning could enhance its performance in these edge cases.

References

[1] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-Pixel Classification is Not All You Need for Semantic Segmentation. In *Advances in Neural Information Processing Systems*, pages 17864–17875. Curran Associates, Inc., 2021. 5

[2] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention Mask Transformer for Universal Image Segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1280–1289, New Orleans, LA, USA, 2022. IEEE. 1, 2, 4, 5

[3] Yong Xien Chng, Henry Zheng, Yizeng Han, Xuchong Qiu, and Gao Huang. Mask Grounding for Referring Image Segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26563–26573, Seattle, WA, USA, 2024. IEEE. 4, 6

[4] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-Language Transformer and Query Generation for Referring Segmentation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16301–16310, Montreal, QC, Canada, 2021. IEEE. 4, 6

[5] Yutao Hu, Qixiong Wang, Wenqi Shao, Enze Xie, Zhenguo Li, Jungong Han, and Ping Luo. Beyond One-to-One: Rethinking the Referring Image Segmentation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4044–4054, Paris, France, 2023. IEEE. 1, 4, 6

[6] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar, 2014. Association for Computational Linguistics. 6

[7] Namyup Kim, Dongwon Kim, Suha Kwak, Cuiling Lan, and Wenjun Zeng. ReSTR: Convolution-free Referring Image Segmentation Using Transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18124–18133, New Orleans, LA, USA, 2022. IEEE. 4

[8] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning Segmentation via Large

Figure 3. **Failure case analysis.** We showcase challenging expressions that lead to incorrect or incomplete segmentation results. Some errors occur due to ambiguous expressions (e.g., "2nd elephant from us"), fine-grained object distinction failures (e.g., "the man in the suit and blue tie behind the old woman's arm"), or complex spatial reasoning (e.g., "a lineup of vehicles situated to the right").

Language Model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9579–9589, Seattle, WA, USA, 2024. IEEE. 1, 4

[9] Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized Referring Expression Segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23592–23601, Vancouver, BC, Canada, 2023. IEEE. 1, 2, 4

[10] Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R. Manmatha. PolyFormer: Referring Image Segmentation as Sequential Polygon Generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18653–18663, Vancouver, BC, Canada, 2023. IEEE. 4

[11] Si Liu, Tianrui Hui, Shaofei Huang, Yunchao Wei, Bo Li, and Guanbin Li. Cross-Modal Progressive Comprehension for Referring Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4761–4775, 2022. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. 6

[12] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019. arXiv:1907.11692. 1, 4

[13] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. 2018. 1

[14] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-Task Collaborative Network for Joint Referring Expression Comprehension and Segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10031–10040, Seattle, WA, USA, 2020. IEEE. 6

[15] Zhuoyan Luo, Yinghao Wu, Cheng Tianheng, Yong Liu, Yicheng Xiao, Wang Hongfa, Xiao-Ping Zhang, and Yujiu Yang.

Cohd: A counting-aware hierarchical decoding framework for generalized referring expression segmentation. *arXiv preprint arXiv:2405.15658*, 2024. 1

[16] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and Comprehension of Unambiguous Object Descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–20, Las Vegas, NV, USA, 2016. IEEE. 6

[17] Hai Nguyen-Truong, E.-Ro Nguyen, Tuan-Anh Vu, Minh-Triet Tran, Binh-Son Hua, and Sai-Kit Yeung. Vision-Aware Text Features in Referring Image Segmentation: From Object Understanding to Context Understanding, 2024. arXiv:2404.08590. 4

[18] Nisarg A. Shah, Vibashan VS, and Vishal M. Patel. LQMFormer: Language-Aware Query Mask Transformer for Referring Image Segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12903–12913, 2024. ISSN: 2575-7075. 4

[19] Chao Shang, Zichen Song, Heqian Qiu, Lanxiao Wang, Fanman Meng, and Hongliang Li. Prompt-Driven Referring Image Segmentation with Instance Contrasting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4124–4134, Seattle, WA, USA, 2024. IEEE. 4

[20] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sébastien Ourselin, and M. Jorge Cardoso. Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations, 2017. arXiv:1707.03237. 5

[21] Jiajin Tang, Ge Zheng, Cheng Shi, and Sibei Yang. Contrastive Grouping with Transformer for Referring Image Segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23570–23580, Vancouver, BC, Canada, 2023. IEEE. 4

[22] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. CRIS: CLIP-Driven Referring Image Segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11676–11685, New Orleans, LA, USA, 2022. IEEE. 4

[23] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. GSVA: Generalized Segmentation via Multimodal Large Language Models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3858–3869, Seattle, WA, USA, 2024. IEEE. 1, 4

[24] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip H.S. Torr. LAVT: Language-Aware Vision Transformer for Referring Image Segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18134–18144, New Orleans, LA, USA, 2022. IEEE. 4, 6

[25] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable Transformers for End-to-End Object Detection, 2021. arXiv:2010.04159. 4