# Supplementary Material for ROSE: Retrieval-Oriented Segmentation Enhancement

Song Tang* Guangquan Jie* Henghui Ding✉ Yu-Gang Jiang
Fudan University, China
https://henghuiding.com/ROSE/

## A. More NEST and NEST+ Dataset details

**NEST Dataset.** NEST Dataset samples are shown in Fig. I. The dataset contains diverse images paired with questions, answers, and masks, generated by an automatic pipeline. To ensure ethical standards, we employ Vision Language Models [7] and trained human reviewers to screen all content, filtering potentially harmful images, questions, or objects with significant negative social impacts. This screening process maintains dataset quality while effectively minimizing risks from visual content.

**NEST+ Dataset.** We construct NEST+ by combining NEST with ReasonSeg [9], RefCOCO [8], RefCOCO+ [8], and RefCOCOg [21]. NEST+ simulates diverse real-world scenarios involving real-time retrieval, reasoning, and referring segmentation. For the NEST split, we randomly sample 500 examples from NEST to represent challenging real-time information retrieval tasks. For the ReasonSeg split, we randomly sampled 200 examples from ReasonSeg's test set to incorporate more complex reasoning scenarios. For the RefSeg split, we randomly sampled 60 examples each from RefCOCO test-A and test-B, 60 examples each from RefCOCO+ test-A and test-B, and 120 examples from RefCOCOg test. In total, RefSeg contains 360 examples. This combination provides a comprehensive benchmark for evaluating models on referring expression segmentation tasks of varying difficulty. Overall, NEST+ contains 1,060 examples across three distinct task categories, providing a challenging evaluation benchmark that comprehensively assesses referring segmentation models.

**Partition Method.** There is no fixed partition in the NEST dataset, as different MLLM-based segmentation models [3, 4] use different foundation MLLM models, which have different knowledge cutoffs and knowledge capacities. To ensure fair comparisons, we maintain consistency by using the same MLLM foundation model (LLaVA-v1.5-7B [11]) for all MLLM-based segmentation baseline methods [9, 14, 19] to partition the NEST dataset into a novel entity split and an emerging entity split. In detail, for each answer $\mathcal{A}$ corresponding to multiple multi-entity images $\mathcal{I}_m$, we use the template "*Please segment* $\{answer\}$ *in this image.*" where $\{answer\}$ is the ground truth answer $\mathcal{A}$ to guide MLLM-based segmentation baseline methods to perform segmentation. For samples where one answer corresponding to a set of multiple images achieves Acc@0.5 exceeding 0.7, we classify all of them into the emerging entity split, while the remaining samples are categorized into the novel entity split. For answer sets with fewer than 3 multi-entity images, we input the answer into the MLLM foundation model and prompt it to return an entity description. We then compare this description with the entity's introduction retrieved from the internet using similarity metrics. If the similarity score falls below a predetermined threshold, all samples in that set are categorized into the novel entity split.

## B. More Implementation Details

Beyond the reproducibility information provided in the main paper, we present additional implementation details to ensure complete transparency and facilitate reproduction. These comprehensive details are outlined in Table I.

Table I. **More Implementation Details.**

| Implementation Detail | Configuration |
|---|---|
| Search engine | DuckDuckGo |
| Image retrieval source | Bing Image Search |
| Object detector | YOLOv8 |
| Number of reference images | $\geq 5$ |
| CLIP similarity threshold | 0.8 |

## C. More experimental results and discussion

**Vanilla Referring Segmentation Results.** To demonstrate that our model is also competent in vanilla referring segmentation tasks [2, 5], we compare it with existing state-of-the-art methods in Table II. We evaluate these methods on the validation and testing sets of RefCOCO, RefCOCO+ [8], and RefCOCOg [21]. While enhancing

*Equal contribution.

✉ Henghui Ding (henghui.ding@gmail.com) is the corresponding author with the Institute of Big Data, College of Computer Science and Artificial Intelligence, Fudan University, Shanghai, China.


Figure I. Data Samples of the NEST Dataset.

Table II. **Referring expression segmentation results** on RefCOCO, RefCOCO+ [8] and RefCOCOg [21] dataset. The cIoU metrics of each split are reported. Baselines are excerpted from [14].

| Method | RefCOCO (UNC) | | | RefCOCO+ (UNC) | | | RefCOCOg (UMD) | |
|---|---|---|---|---|---|---|---|---|
| | Val. | Test-A | Test-B | Val. | Test-A | Test-B | Val. | Test |
| MCN [12] | 62.4 | 64.2 | 59.7 | 50.6 | 55.0 | 44.7 | 49.2 | 49.4 |
| VLT [1] | 67.5 | 70.5 | 65.2 | 56.3 | 61.0 | 50.1 | 55.0 | 57.7 |
| CRIS [18] | 70.5 | 73.2 | 66.1 | 62.3 | 68.1 | 53.7 | 59.9 | 60.4 |
| ReLA [10] | 73.8 | 76.5 | 70.2 | 66.0 | 71.0 | 57.7 | 65.0 | 66.0 |
| SEEM [23] | - | - | - | - | - | - | 65.7 | - |
| LISA-7B [9] | 74.9 | 79.1 | 72.3 | 65.1 | 70.8 | 58.1 | 67.9 | 70.6 |
| SESAME-7B [19] | 74.7 | - | - | 64.9 | - | - | 66.1 | - |
| READ-7B [14] | 78.1 | 80.2 | 73.2 | **68.4** | **73.7** | **60.4** | 70.1 | **71.4** |
| LISA-7B + ROSE (ours) | 76.2 | 80.1 | 73.5 | 66.3 | 71.1 | 58.9 | 65.4 | 68.6 |
| SESAME-7B + ROSE (ours) | 76.1 | 79.7 | 73.3 | 64.8 | 69.6 | 56.8 | 67.3 | 70.2 |
| READ-7B + ROSE (ours) | **79.3** | **80.8** | **74.0** | 66.2 | 72.5 | 60.2 | **70.3** | 70.2 |

Table III. **Reasoning Segmentation results.** Result comparison on the ReasonSeg dataset. Baselines are excerpted from [14].

| Method | Validation Set | | Test Set | |
|---|---|---|---|---|
| | gIoU | cIoU | gIoU | cIoU |
| X-Decoder [22] | 22.6 | 17.9 | 21.7 | 16.3 |
| SEEM [23] | 25.5 | 21.2 | 24.3 | 18.7 |
| GRES [10] | 22.4 | 19.9 | 21.3 | 22.0 |
| LISA-7B [9] | 52.9 | 54.0 | 47.3 | 48.4 |
| SESAME-7B [19] | 34.8 | 39.1 | 30.5 | 30.4 |
| READ-7B [14] | **59.8** | **67.6** | **56.8** | **59.0** |
| LISA-7B + ROSE (ours) | 51.7 | 52.6 | 46.3 | 47.4 |
| SESAME-7B + ROSE (ours) | 32.1 | 38.2 | 29.3 | 27.9 |
| READ-7B + ROSE (ours) | 57.6 | 65.2 | 53.1 | 56.6 |

novel emerging segmentation capabilities, our model also achieves comparable results across these various referring segmentation methods.

**Reasoning Segmentation Results.** We further evaluate our model's capability in reasoning segmentation tasks by comparing it against current leading methods in Table III. The evaluation focuses on datasets that require complex reasoning to interpret implicit instructions before performing segmentation. Results demonstrate that ROSE achieves comparable results on reasoning segmentation.

**Discussion on SESAME and READ's Experimental Results on the NEST Dataset.** On the NEST dataset, SESAME-7B [19] (13.1% in gIoU) and READ-7B [14] (22.5% in gIoU) experience significant performance drops on the NEST dataset. This is because they follow a strategy of first correcting wrong referents and then adjusting the input prompt with an alternative to segment the closest object when encountering an empty target. However, for the NEST dataset, their MLLM foundation models lack knowledge about the novel and emerging entities in the questions, resulting in mostly empty targets and consequently incorrect adjusted input prompts. On the one hand, this demonstrates the challenging nature of the NEST dataset, which goes beyond existing MLLM-based segmentation methods' knowledge cutoff. On the other hand, this limitation highlights the importance of retrieval-augmented generation (RAG) methods for handling novel emerging segmentation.

## D. More Ablation Studies

Table IV. **Vision Encoder Selection in VPE.**

| Vision Encoder | gIoU | cIoU |
|---|---|---|
| ResNet-50 [6] | 63.4 | 59.3 |
| VGG-16 [17] | 59.3 | 55.2 |
| DINOv2-ViT-L/14 [13] | 65.0 | 61.4 |
| **CLIP-ViT-L/32** [15] | **73.0** | **68.6** |

**Vision Encoder Selection in VPE.** We evaluate different pre-trained vision encoders for our Visual Prompt Enhancer (VPE) module on the NEST dataset. As shown in Table IV,

Table V. **Runtime Performance Comparison.**

| Method | RAG | Time per Image (s) | FPS (frames/s) | gIoU |
|---|---|---|---|---|
| LISA-7B | ✗ | 4.67 | 0.21 | 48.7 |
| LISA-7B+ GPT-4o mini Search | ✓ | 9.12 | 0.11 | 53.5 |
| LISA-7B+ Gemini-2.0 Flash Search | ✓ | 8.13 | 0.12 | 53.8 |
| LISA-7B+ ROSE (ours) | ✓ | 9.67 | 0.10 | 73.0 |

CLIP-ViT-L/32 [15] delivers the best performance, achieving 73.0 gIoU and 68.6 cIoU, significantly outperforming other vision encoders. Based on these results, we adopt CLIP-ViT-L/32 as our design choice for VPE module.

**Latency and Computational Efficiency.** We acknowledge that incorporating the RAG module does introduce additional computational overhead compared to standard inference pipelines. However, ROSE achieves significant and consistent performance improvements while maintaining comparable runtime costs to commercial RAG models through a carefully designed and optimized system architecture. As shown in Table V, our method achieves a substantial 19.5 % improvement compared to GPT-4o mini Search (73.0 vs 53.5), while only adding approximately 0.55 seconds of additional latency to the total runtime (9.67s vs 9.12s), demonstrating a highly favorable performance-efficiency tradeoff. This efficiency is achieved through targeted system optimizations including parallelization strategies (e.g., asynchronously downloading retrieved images while the MLLM is simultaneously processing the query), which effectively minimizes idle waiting time across modules. The detailed runtime breakdown for each individual module is presented in Table VI, where values denote average processing time in seconds per sample. All experiments are conducted on the NEST dataset under identical and controlled network conditions to ensure fair comparison.

Table VI. **Module Processing Time Breakdown.**

| Module | Processing Time (s) |
|---|---|
| WebSense | 0.35 |
| IRAG | 6.27 |
| TPE | 0.83 |
| VPE | 1.68 |

**Necessity of MLLM.** To validate the necessity of incorporating MLLM alongside VPE, we compare a simple IRAG+VPE pipeline against our full ROSE framework across multiple benchmark splits. As shown in Table VII, while the simple IRAG+VPE pipeline achieves reasonable performance on NEST (68.4%), it shows significant and consistent limitations in more complex scenarios like ReasonSeg (13.1%) and RefSeg (20.9%), where deeper semantic understanding is required. This performance degradation occurs because the simple pipeline indiscriminately sends all queries to internet retrieval, returning irrelevant or noisy images for cases that do not require external knowledge, thereby introducing harmful distractions. Therefore, the MLLM component is essential as it provides strong reasoning capabilities and serves as a robust fallback mechanism when retrieval is unnecessary or uninformative. Our full ROSE framework effectively addresses these limitations by intelligently and adaptively determining when retrieval is truly necessary, leading to more accurate and reliable segmentation across diverse task categories.

Table VII. **The Necessity of MLLM Component.**

| Method | NEST | ReasonSeg | RefSeg | Overall |
|---|---|---|---|---|
| LISA-7B | 51.1 | **42.5** | **54.9** | 50.9 |
| IRAG+VPE | 68.4 | 13.1 | 20.9 | 53.1 |
| ROSE | **75.3** | 42.2 | 54.4 | **67.6** |

## E. Limitations

There are certain limitations in our NEST dataset that are worth acknowledging for future improvement. The average number of effective visual entities per image is relatively low (2.7), which means that some models can achieve apparently correct segmentation through hallucination rather than genuine understanding. Traditional segmentation methods [16, 18, 20, 23] without MLLMs and LISA-7B [9] achieved gIoU scores between 40% and 50% on the NEST dataset, which initially appear competitive. Upon closer analysis, however, we found that these methods tend to segment the main entities or centrally located entities in the image, regardless of the specific query or its semantic content. This shortcut behavior can lead to inflated and misleading performance metrics when the target entity happens to be the dominant or central object in the image, obscuring the true reasoning capability of the model. Future iterations of the dataset could therefore benefit from including more complex and diverse scenes with a higher number of competing visual entities to more rigorously evaluate model precision. In parallel, we will also optimize our data engine to produce more challenging and varied data samples that better expose the limitations of models relying on positional or saliency-based shortcuts.

## F. More Visualizations

We provide additional qualitative visualizations in Fig. II to further demonstrate the effectiveness and robustness of our ROSE method in handling diverse, novel, and emerging entity segmentation across a variety of real-world scenarios.

Please segment Apple Vision Pro.

ROSE

Sure, [SEG]

Please segment PlayStation 5.

ROSE

Sure, it is [SEG]

Which character from The Walking Dead will appear as an NPC in Fortnite starting April 8, 2025?

ROSE

Sure, the segmentation result is [SEG]

Who launched a global tour covering five continents in 2024 that set ticket sale records?

ROSE

Sure, [SEG]

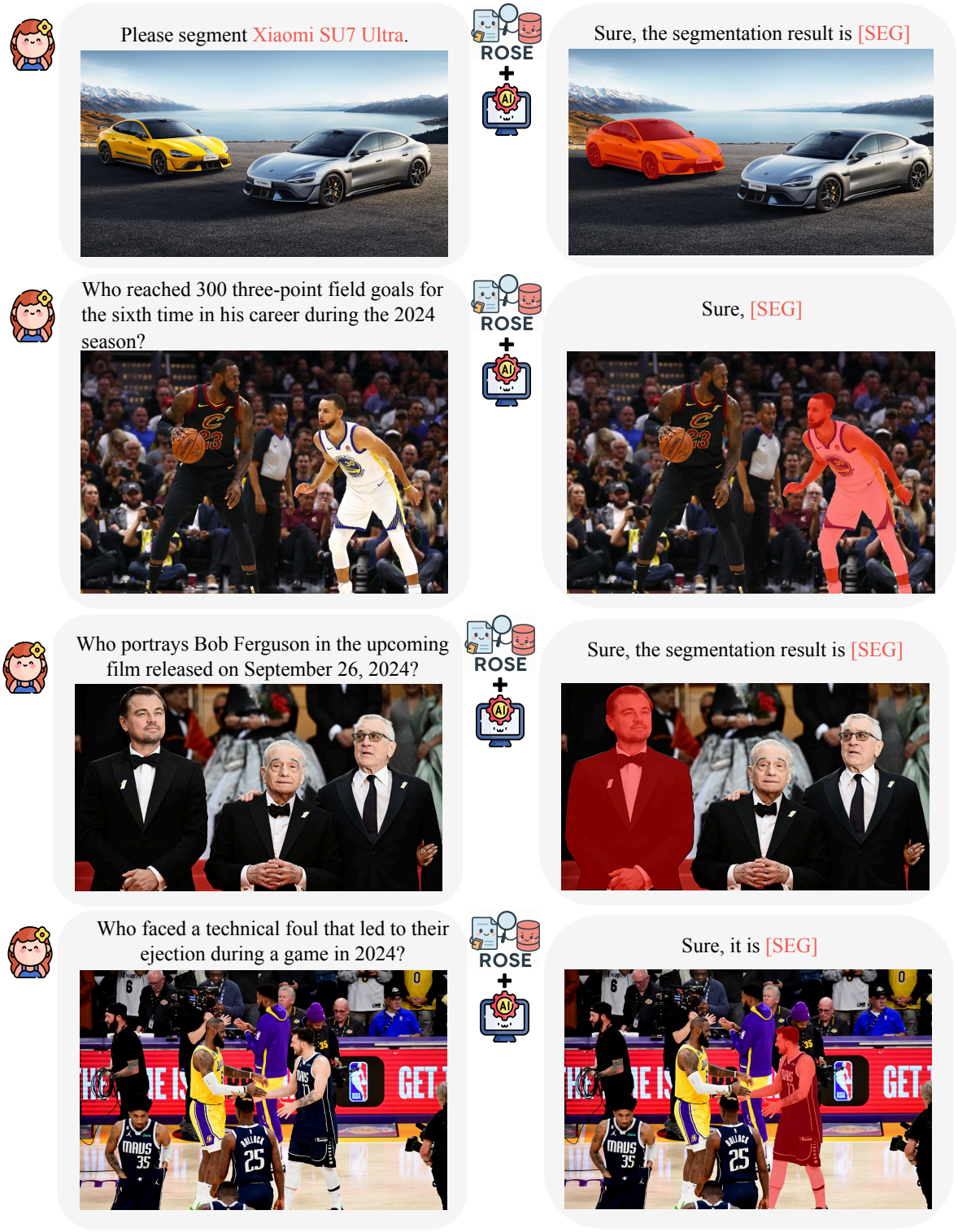


Figure II. **More visualizations.** ROSE demonstrates superior performance on both emerging and novel entity segmentation.

## References


[1] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *ICCV*, 2021. 2

[2] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. VLT: Vision-language transformer and query generation for referring segmentation. *IEEE TPAMI*, 45(6), 2023. 1

[3] Henghui Ding, Chang Liu, Shuting He, Kaining Ying, Xudong Jiang, Chen Change Loy, and Yu-Gang Jiang. MeViS: A multi-modal dataset for referring motion expression video segmentation. *IEEE TPAMI*, 2025. 1

[4] Henghui Ding, Song Tang, Shuting He, Chang Liu, Zuxuan Wu, and Yu-Gang Jiang. Multimodal referring segmentation: A survey. *arXiv preprint arXiv:2508.00265*, 2025. 1

[5] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Yu-Gang Jiang. GREx: Generalized referring expression segmentation, comprehension, and generation. *IJCV*, 2026. 1

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[7] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1

[8] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 1, 2

[9] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *CVPR*, 2024. 1, 2, 3

[10] Chang Liu, Henghui Ding, and Xudong Jiang. GRES: generalized referring expression segmentation. In *CVPR*, 2023. 2

[11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2024. 1

[12] 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 *CVPR*, 2020. 2

[13] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[14] Rui Qian, Xin Yin, and Dejing Dou. Reasoning to attend: Try to understand how <SEG> token works. In *CVPR*, 2025. 1, 2

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3

[16] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 3

[17] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[18] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *CVPR*, 2022. 2, 3

[19] Tsung-Han Wu, Giscard Biamby, David Chan, Lisa Dunlap, Ritwik Gupta, Xudong Wang, Joseph E Gonzalez, and Trevor Darrell. See say and segment: Teaching lmms to overcome false premises. In *CVPR*, 2024. 1, 2

[20] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *CVPR*, 2024. 3

[21] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016. 1, 2

[22] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *CVPR*, 2023. 2

[23] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *NeurIPS*, 36, 2024. 2, 3