MaskRAG: Mask Retrieval Augmented Generation for MLLM-based Referring Expression Segmentation

Zhongjiang He[1,2], An Zhao[2*], Canhui Tang[3,2], Hao Sun[2], Hongbo Sun[2], Ye Yuan[2], Kongming Liang[1,4], Zhanyu Ma[1,4*]
[1]Beijing University of Posts and Telecommunications, Beijing, China
[2]China Telecom Artificial Intelligence Technology (Beijing) Co., Ltd, Beijing, China
[3]Xi'an Jiaotong University, Xi'an, China
[4]Beijing Key Laboratory of Multimodal Data Intelligent Perception and Governance

***Abstract*—Multimodal Large Language Models (MLLMs) have significantly advanced referring expression segmentation (RES) by understanding complex language instructions. Current approaches typically employ a `[SEG]` token as a textual prompt to bridge the MLLM and downstream segmentation models (*e.g.*, SAM). However, due to the weak prompt nature of the `[SEG]` token, existing frameworks often suffer from the "segmentation hallucination" issue, where the model confidently predicts structures or objects that are absent from ground truth. In this paper, we propose MaskRAG, a multimodal RAG-like framework that advances RES through adaptive mask retrieval and augmentation mechanisms. Specifically, we propose a mask retrieval module that efficiently encodes region features and embeds those features with a customized language template, enabling finer-grained perception of scenes and discrimination among candidate regions. Furthermore, we propose a mask augmentation module with multi-granularity semantic fusion and adaptive routing mechanisms, which effectively leverages the strengths of both `[SEG]`-based segmentation and our retrieval approaches. Experiments demonstrate that MaskRAG achieves state-of-the-art performance across multiple benchmarks.**

***Index Terms*—Multimodal Large Language Models, Referring expression segmentation**

I. Introduction

Referring Expression Segmentation (RES) [1] is a multimodal task that aims to segment objects in an image based on natural language descriptions, with applications in robotics manipulation and image editing. Multimodal Large Language Models (MLLMs) [2]–[4] are foundation models that process and align information across visual and textual modalities through joint embedding spaces, enabling visual grounding and language-guided reasoning about images. The introduction of MLLMs has revolutionized RES, enabling it to interpret complex language commands [5], [6].

Recent advances [5], [7], [8] adopt a `[SEG]`-based paradigm. The `[SEG]` token serves as a textual prompt to bridge MLLMs and downstream segmentation models, *e.g.*, SAM [9] or SAM2 [10]. However, with the `[SEG]` token being the only interface between the language and segmentation models, the prompt capacity is weak. Therefore, the output mask is often deficient. For instance, the model may over-segment irrelevant regions, mislocalize the target object, or yield fragmented segmentation results (Fig. 1). This can be viewed as a form of hallucination, as the model confidently predicts structures or objects that are absent from ground truth—a phenomenon we refer to as "*segmentation hallucination*".

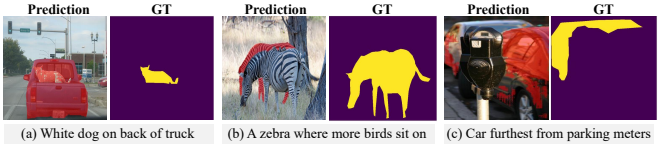

Fig. 1. Common issues in `[SEG]`-based methods: (a) over-segmentation; (b) mislocalization of target object; (c) fragmented segmentation results.

Inspired by Retrieval Augmented Generation (RAG) [11], [12] in mitigating the hallucination issues of language generation of LLMs, we ask: *can we propose a RAG-like multimodal segmentation framework to address segmentation hallucination?* However, achieving this objective is non-trivial due to the inherent gap between the segmentation task and the language generation task of LLMs. This discrepancy necessitates a rethinking of both the "retrieval" and "augmentation" concepts within the context of RES task.

In this paper, we propose **MaskRAG** (Fig. 2), a novel framework that advances RES through adaptive mask retrieval and augmentation mechanisms. While "retrieval" traditionally means searching databases, our method embodies retrieve-then-augment: LLM-guided mask selection from proposals (retrieval), then augmentation of decoding. This adapts RAG's core principle to vision for addressing multimodal hallucination. First, we introduce a **mask retrieval** module, an innovative retrieval-based alternative to the conventional paradigm that uses a segmentation decoder to regress masks. Our

*Corresponding authors: Zhanyu Ma (mazhanyu@bupt.edu.cn) and An Zhao (zhaoa@chinatelecom.cn). This work was supported in part by the National Natural Science Foundation of China (Grant 62225601, U23B2052), in part by the Beijing Natural Science Foundation Project No. L242025, and in part by the Beijing Key Laboratory of Multimodal Data Intelligent Perception and Governance.

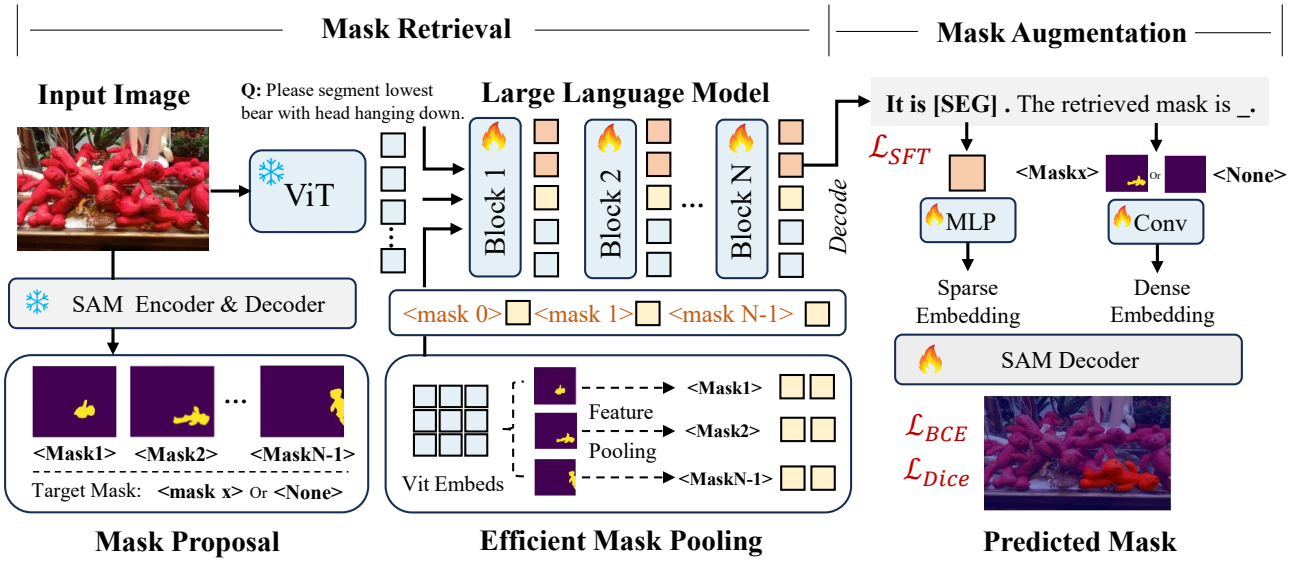

Fig. 2. Overview of the MaskRAG framework. In the Mask Retrieval module, mask proposals undergo visual feature extraction through efficient mask pooling and are integrated with a specialized language template for the LLM to predict the best mask. In the Mask Augmentation module, mask prompts are adaptively projected into dense embeddings to improve SAM2 decoding.

framework first uses SAM2 to generate a candidate mask set, each capturing potential object instances in the scene. These candidates undergo visual feature extraction via an efficient mask pooling operation and are combined with specialized language embeddings. Subsequently, the language model is taught to identify and retrieve the optimal mask index through instruction tuning.

Second, we propose an **mask augmentation** module that integrates the strengths of both `[SEG]`-based and our retrieval-based approach. Specifically, the `[SEG]` token is encoded into a sparse embedding, while the mask prompt is transformed into a dense embedding, jointly activating SAM2's segmentation capability. Although the retrieved mask offers notable advantages, its performance is limited by the quality of candidate proposals. To mitigate this, we introduce a `<None>`-token-based routing mechanism that allows dynamic choice between using retrieved masks or relying solely on the `[SEG]` token. Furthermore, we concatenate hidden states from different LLM layers to incorporate multigranularity semantics, improving segmentation accuracy.

The MaskRAG framework mitigates segmentation hallucination through two key advantages: (1) the design of the proposal-then-retrieval paradigm enables finer-grained perception and discrimination among candidate image regions, and (2) the use of high-quality masks from raw SAM2 as robust prompts, which guide the on-the-fly tuning SAM2 model to maintain structurally sound segmentations, reducing fragmented outputs. Comprehensive experiments demonstrate that MaskRAG achieves state-of-the-art performance across multiple benchmarks. Our contributions are as follows:

- We investigate the "segmentation hallucination" issue in the RES task, and propose a multimodal RAG-like referring segmentation framework to address it.
- We propose a mask retrieval module that efficiently encodes region features and embeds them with a customized language template, enabling finer-grained perception of scenes and discrimination among candidates.
- We propose a mask augmentation module with multi-granularity semantic fusion and adaptive routing mechanisms, which effectively leverages the strengths of both `[SEG]`-based and our retrieval-based approaches.

The code of MaskRAG is available at https://github.com/telecomhzj/MaskRAG.

II. Related Work

A. MLLM

Building upon the general understanding capabilities of LLMs [13], [14], significant research efforts have been devoted to advancing MLLMs. Prevalent works, such as BLIP-2 [3], LLaVA [2], and Qwen2.5VL [4], have narrowed the gap between image and text modalities through visual instruction tuning, enabling the development of multimodal assistant models capable of visual perception and visual question answering. Researchers have also introduced numerous studies aimed at enhancing visual understanding capabilities [15]–[18] and video understanding capabilities [19]–[21]. Despite these efforts, while MLLMs excel at visual-semantic reasoning, they lack the precise perception capabilities required for dense prediction tasks, necessitating integration with specialized architectures like SAM [9].

B. MLLMs for RES

Compared to classic semantic segmentation, RES introduces dual challenges: understanding complex natural language expressions and achieving fine-grained visual understanding of

objects, including their attributes, spatial relationships, and interactions within the scene. RES methods are typically two-stage or one-stage. Two-stage methods generate region proposals using segmentation models, then select the correct region based on the description. One-stage methods fuse visual and textual features to directly predict the target mask [22]. With the advent of MLLMs, recent RES approaches exploit their superior visual-language reasoning to achieve more accurate segmentation. LISA [5] pioneered the `[SEG]` token paradigm to bridge MLLMs and downstream segmentation models (e.g., SAM [9] or SAM2 [10]). Subsequently, PixelLM [7] employs multiple `[SEG]` tokens for multi-object mask segmentation. SAM4MLLM [23] proposes a two-step approach, where MLLMs first predict bounding boxes and then determine whether a point lies inside an object to select optimal point prompts for SAM. Sa2VA [8] combines InternVL-2.5 [24] and SAM2 to develop a versatile system supporting image and video referring segmentation. However, prior works overlook the segmentation hallucination issue of `[SEG]` tokens, which our work mitigates by proposing MaskRAG.

III. Method

The entire pipeline of our MaskRAG framework is illustrated in Fig. 2, which consists of a mask retrieval module that predicts the best proposal index and an adaptive mask augmentation module that enhances segmentation decoding. The pseudocode is presented in Algorithm 1.

Algorithm 1: Pseudo-code of MaskRAG

Input: $\mathbf{x}_{img}$, $\mathbf{x}_{txt}$
Output: ouput text $\mathbf{o}_{txt}$, predicted mask $\hat{\mathbf{M}}$
/* **(1) Mask Retrieval** */
$\mathbf{M}_p \leftarrow$ SAM2 AutoGenerator; # Mask proposal
$\mathbf{F}_v, \mathbf{F}_s \leftarrow$ Vision encoder, SAM2 encoder ;
$\mathbf{F}_p \leftarrow$ Pooling ($\mathbf{F}_v$, resize($\mathbf{M}_p$)) # Mask embedding;
$\mathbf{M}_R, \mathbf{F}_h, \mathbf{o}_{txt} \leftarrow$ LLM ($\mathbf{F}_v$, $\mathbf{F}_p$, $\mathbf{x}_{txt}$) # Retrieval;
/* **(2) Mask Augmentation** */
$\mathbf{F}_{sp} \leftarrow$ MLP (concat($\mathbf{F}_c^{1:L}$)); # sparse embedding
if *<None> in* $\mathbf{o}_{out}$ **then**
  $\hat{\mathbf{M}} \leftarrow \mathrm{SAM2}_{dec}$ ($\mathbf{F}_s$, $\mathbf{F}_{sp}$);
else
  $\mathbf{F}_d \leftarrow \mathcal{C}$ ($\mathbf{M}_R$); # dense embedding
  $\hat{\mathbf{M}} \leftarrow \mathrm{SAM2}_{dec}$ ($\mathbf{F}_s$, $\mathbf{F}_{sp}$, $\mathbf{F}_d$);
end

A. Revisiting *[SEG]*

The `SEG`-based framework was first introduced in LISA [5]. It trains an MLLM to take an image $\mathbf{x}_{img}$ and a text input $\mathbf{x}_{txt}$, predict a `[SEG]` token, and extract the corresponding last hidden state feature $\mathbf{F}_h$, formulated as follows:

$$\mathbf{F}_h, \mathbf{o}_{txt} = \mathrm{MLLM}\left(\mathbf{x}_{img}, \mathbf{x}_{txt}\right), \tag{1}$$

where $\mathbf{o}_{txt}$ denoted the output text of LLM. Additionally, LISA utilizes the SAM encoder to obtain visual image features $\mathbf{F}_s$, and projects $\mathbf{F}_h$ to a sparse embedding $\mathbf{F}_{sp}$ as a prompt to decode the predicted mask $\hat{\mathbf{M}}$.

$$\hat{\mathbf{M}} = \mathrm{SAM}_{dec}\left(\mathbf{F}_s, \mathbf{F}_{sp}\right). \tag{2}$$

B. Mask Retrieval Module

This module transforms conventional mask regression into a retrieval-based framework through mask proposal generation, mask embedding, and LLM-based retrieval.

Mask proposal. For an input image $\mathbf{x}_{img}$, a set of mask proposals is generated via SAM2 [10]'s automatic mask generation. Using uniformly sampled point prompts, SAM2 produces N high-quality proposals $\mathbf{M_p} \in \mathbb{R}^{N\times H\times W}$, each capturing potential object instances. These proposals are designed to segment arbitrary objects in the image. After post-processing like IoU filtering, about 10$\sim$50 proposals remain.

Efficient mask embedding. To efficiently encode these mask proposals, we directly reuse the visual embeddings from the MLLM. The input image is encoded into image embeddings $\mathbf{F}_v \in \mathbb{R}^{H'\times W'\times C}$ using the MLLM's vision encoder. With the resized mask proposals $\mathbf{M'_p} \in \mathbb{R}^{N\times H'\times W'}$, the visual embeddings are partitioned into N regional features. For each region, we uniformly pool the features into S tokens, resulting in mask embeddings $\mathbf{F}_p \in \mathbb{R}^{N\times S\times C}$.

LLM-based mask retrieval. As shown in Fig. 2, each mask proposal $\mathbf{F}_p^i$ is prefixed with a special token `<mask i>`. These tokens are then concatenated with the original image tokens to form the input sequence. The text ground truth is determined by computing the IoU between each proposal and the ground truth mask. If the maximum IoU exceeds a threshold m, the proposal with the highest IoU is selected as the ground truth, and the answer is formatted as "The retrieved mask is `<mask x>`". Otherwise, if no proposal has an IoU above m, the answer is set to "The retrieved mask is `<None>`". The MLLM is trained via supervised fine-tuning using these designed responses. Using the predicted mask index, we can get the corresponding mask proposal $\mathbf{M}_R$.

C. Mask Retrieval Augmentation Generation

This module utilizes the retrieved mask to adaptively enhance the `[SEG]`-based segmentation mask decoding process.

Multi-granularity semantic fusion. To generate the final mask, the retrieved mask $\mathbf{M}_R$ acts as a dense prompt $\mathbf{F}_d \in \mathbb{R}^{H_d\times W_d\times D}$ through a convolutional neural network $\mathcal{C}$, and the hidden state of the `[SEG]` token serves as a sparse prompt $\mathbf{F}_{sp} \in \mathbb{R}^{1\times D}$ via an MLP, jointly activating SAM2's segmentation capability. To enhance the representational capacity of the `[SEG]` token, we utilize the hidden states from selected L layers of `[SEG]` rather than only the last layer to generate a sparse prompt. Specifically, the concatenated feature $\mathbf{F}_c \in \mathbb{R}^{L\times D}$ is then projected to $\mathbf{F}_{sp} \in \mathbb{R}^{1\times D}$.

Adaptive routing mechanism. Although the retrieved mask offers notable advantages, we observe that its performance is inherently constrained by the quality of the candidate proposals, primarily due to inevitable missed detections. To address this, we introduce a `<None>`-token-based routing mechanism. When the retrieved candidates are deemed inadequate, the

model learns to output the <None> token and switch back to the original [SEG]-only prediction branch.

Training loss. Following LISA [5], the overall training objective includes a next-token prediction loss to supervise the generated mask index token and [SEG] token, as well as a binary cross-entropy (BCE) loss and Dice loss computed between the predicted mask $\hat{\mathbf{M}}$ and the ground truth mask.

IV. Experiments

A. Experimental Setup

Datasets and metrics. Following prior studies [30], we train our model on the RefCOCO series [33], [34] for RES task. Unlike existing methods [5], [25] that rely on mixed-dataset pretraining before fine-tuning on RefCOCO series, our approach attains state-of-the-art results with training solely on the RefCOCO/+/g training set. For the metric, our evaluation uses the widely recognized class-wise IoU (cIoU) metric.

Implementation details. We employ InterVL2.5-8B [24] and SAM2 [10] as the backbone. The LLM is finetuned using LoRA [35] with a rank of 128 and a learning rate of 4×10^{-5}. The MLP and SAM2 are trained with the same learning rate. Training requires 20 hours on 8 NVIDIA A800 GPUs. The max number of mask proposals is 50, the IoU threshold m is 0.8, and the number of pooled tokens S is 10. H' and W' are both 16, and H_d and W_d are both 256. The multi-layer hidden states are taken from the 1/4, 2/4, 3/4, and final layers. Besides the [SEG] token, we introduce <mask1>-<mask50> to index each proposal, along with <None> to extend and standardize the output.

Baseline. We construct our baseline using only [SEG] token. In this setting, the segmentation mask is generated solely using the hidden state of the [SEG] token as the prompt to SAM2, without employing mask retrieval or the proposed MaskRAG mechanism. We also evaluated our framework against other LLM-based approaches across RefCOCO, RefCOCO+ and RefCOCOg validation and test samples.

B. Quantitative Comparison

Table I presents the comparison results of our approach against state-of-the-art MLLM-based RES models. Across all benchmarks, our method consistently demonstrates significant performance improvements on RefCOCO, RefCOCO+, and RefCOCOg. Specifically, it surpasses SegAgent [31] by 3.6 and UFO [29] by 4.4. We also provide a 4B model with a smaller LLM backbone. Despite its reduced size, this model only experiences a marginal performance drop of 1.2 compared to our 8B version, while still outperforming SegAgent by 2.4 in cIoU.

To isolate our framework's contribution from backbone capacity, we conduct two types of controlled comparisons: (1) Our 4B model outperforms all 7B baselines, demonstrating that MaskRAG's superior performance derives from our method's design rather than requiring larger model scale. (2) When directly matched against MLLMSeg using the same InternVL2.5-8B backbone (82.1 vs. 81.0), our method maintains a clear advantage. Together, these comparisons confirm that performance gains originate from our framework design—effective even with smaller backbones and consistent when backbone capacity is held constant.

C. Qualitative Comparison

We present visualized comparisons with LISA [5], the baseline, and ours. The red regions indicate predicted masks, and the baseline integrates InternVL2.5 and SAM2 using only the [SEG] token. As shown in Fig. 3, LISA and the baseline fail to accurately locate the "yellow tennis" and the "car furthest parking meter", while our method succeeds. These results demonstrate our approach effectively mitigates segmentation hallucination issues.

Importantly, while cIoU improvements appear modest on saturated benchmarks (Table I), this metric has inherent limitations in capturing structural correctness. Fragmented or incomplete masks can achieve reasonable cIoU scores despite being unusable in practice. Our method prioritizes structural integrity and semantic accuracy, with these advantages primarily evident in the qualitative improvements shown in Fig. 3.

D. Ablation Studies

We conduct an ablation study on MaskRAG's core components (Table II). Row (b) establishes retrieval's performance ceiling: when proposals succeed, retrieval substantially improves perception quality. Row (c) shows the baseline using only [SEG], which achieves suboptimal results due to segmentation hallucinations. Row (d) evaluates retrieval-only methods that directly return retrieved masks, revealing that retrieval alone is limited by the quality of candidate proposals. Simply combining these masks with [SEG] yields unsatisfactory performance (Row (e)), as naive fusion indiscriminately uses all masks and fails when proposals miss targets. Introducing the <None> token substantially improves results (Row (g)) by enabling the model to assess proposal reliability and selectively utilize retrievals. Critically, <None> appears in only 20% of cases—retrieval succeeds 80% of the time. This high retrieval success rate explains our model's strong performance: in the majority of cases, retrieval provides superior masks for mask augmentation module, while adaptive routing handles the minority of cases where proposals fail to capture the referring expression. Finally, incorporating multi-layer concatenation further enhances performance (Row (h)).

E. Computational Complexity Analysis

We report the runtime, memory footprint, and FLOPs of our method compared to the baseline in Table III. While our approach incurs additional overhead (inference time increases from 0.61 s to 1.36 s, memory from 22.39 GB to 25.55 GB, and GFLOPs from 48.24 to 56.48), this cost effectively addresses segmentation hallucination and significantly improves qualitative mask quality. In particular, our method eliminates fragmentation and over-segmentation, producing structurally coherent masks that are suitable for downstream applications.

The additional runtime overhead is mainly due to the SAM2 proposal stage, which takes approximately 0.6 s per image.

TABLE I
COMPARISON RESULTS (CIOU) OF MLLM-BASED REFERRING EXPRESSION SEGMENTATION ON REFCOCO/+/G DATASETS.

Methods	Size	Venue	RefCOCO			RefCOCO+			RefCOCOg		Avg.
			Val	TestA	TestB	Val	TestA	TestB	Val	Test	
LISA [5]	7B	CVPR'24	74.1	76.5	71.1	62.4	67.4	56.5	66.4	68.4	67.9
LISA(ft) [5]	7B	CVPR'24	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6	69.9
GSVA(ft) [25]	7B	CVPR'24	77.2	78.9	73.5	65.9	69.6	59.8	72.7	73.3	71.4
PixelLM [7]	7B	CVPR'24	73.0	76.5	68.2	66.3	71.7	58.3	69.3	70.5	69.2
GLaMM [6]	7B	CVPR'24	79.5	83.2	76.9	72.6	78.7	64.6	74.2	74.9	75.6
VisionLLM v2 [26]	7B	NIPS'24	76.6	79.3	74.3	64.5	69.8	61.5	70.7	71.2	71.0
M²SA [27]	7B	ICLR'25	74.0	76.8	69.7	63.1	67.2	56.1	67.0	68.3	67.8
VRS-HQ [28]	7B	CVPR'25	73.5	77.5	69.5	61.7	67.6	54.3	66.7	67.5	67.3
UFO [29]	8B	arXiv'25	78.0	79.7	75.6	72.3	76.8	66.6	73.7	74.3	74.6
Text4Seg [30]	7B	ICLR'25	79.2	81.7	75.6	72.8	77.9	66.5	74.0	75.3	75.4
SegAgent [31]	7B	CVPR'25	79.7	81.4	76.6	72.5	75.8	66.9	75.1	75.2	75.4
MLLMSeg [32]	8B	arXiv'25	79.2	81.0	77.5	73.9	77.6	68.6	77.3	78.6	76.7
MaskRAG (Ours)	4B	-	80.7	83.2	77.7	75.8	79.8	69.8	77.6	77.9	77.8
MaskRAG (Ours)	8B	-	**82.1**	**83.7**	**78.8**	**77.0**	**80.9**	**72.2**	**78.3**	**78.8**	**79.0**

TABLE II
ABLATION STUDY OF OUR MASKRAG. THE "PERFORMANCE" DENOTES CIOU ON THE VAL SPLIT OF REFCOCO/+/G. GT: GROUND TRUTH. MC: MULTI-LAYER CONCATENATION.

Experiments	Performance
(a) Retrieved by GT	79.0 / 79.5 / 75.8
(b) [SEG] + Retrieved by GT + `[None]` + MC	84.8 / 83.6 / 82.5
(c) `[SEG]` (Baseline)	81.1 / 75.8 / 77.5
(d) Retrieved Mask	72.3 / 68.1 / 67.4
(e) `[SEG]` + Retrieved Mask	80.1 / 74.4 / 75.8
(g) `[SEG]` + Retrieved Mask + `[None]`	81.7 / 76.4 / 78.0
(h) `[SEG]` + Retrieved Mask + `[None]` + MC (Ours)	**82.1 / 77.0 / 78.3**

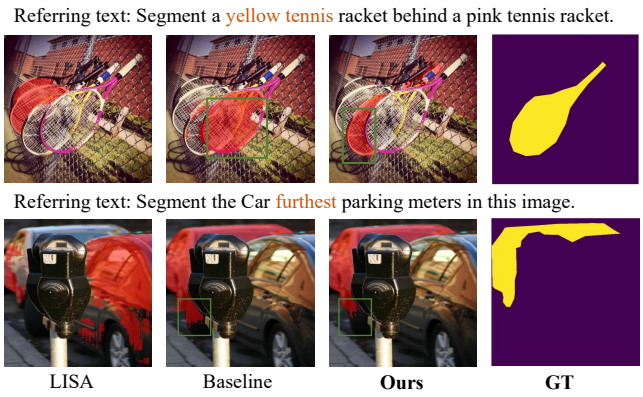

Fig. 3. Qualitative comparisons with existing methods.

This cost is incurred only once per image and can be reduced through faster proposal methods, while the resulting proposals can be reused for multiple objects. Therefore, the overall efficiency remains acceptable for applications that require structurally reliable masks.

TABLE III
INFERENCE TIME, MEMORY USAGE, AND COMPUTATIONAL COST.

Method	Inference time (s)	Memory usage (GB)	GFLOPs
Baseline	0.61	22.39	56.48
Ours	1.36	25.55	48.24

V. DISCUSSION AND CONCLUSION

In this paper, we propose MaskRAG, a novel RES framework that improves segmentation accuracy through adaptive mask retrieval and augmentation. By actively searching for the most relevant mask proposals to guide segmentation decoding, MaskRAG enhances both robustness and precision. Beyond performance gains, our framework provides insights into how multimodal RAG-style systems can be designed to mitigate multimodal hallucination. Extensive experiments demonstrate that MaskRAG consistently achieves state-of-the-art performance on referring expression segmentation benchmarks.

This work focuses on MLLM-based referring segmentation, and experiments are therefore conducted primarily on segmentation-related tasks. Although MaskRAG is developed for RES (similar to LISA [5], SegAgent [31], and MLLM-Seg [32]), it introduces design principles that may generalize to other vision-language tasks, including (1) retrieval-based augmentation to mitigate visual hallucinations, (2) adaptive routing mechanisms for uncertainty handling, and (3) multi-granularity fusion for enhancing performance. MaskRAG is also compatible with general multimodal learning: LoRA preserves the base model parameters, SAM2 operates as an independent module, and prior studies indicate that joint training with VQA tasks can maintain multi-task performance. This suggests that task-focused research may yield reusable mechanisms that improve multimodal reliability more broadly. A broader evaluation on general vision-language tasks is left for future work.

Despite its effectiveness, MaskRAG has two limitations. First, candidate region generation depends on external models such as SAM, which increases inference latency and computational cost. Second, in complex scenes (e.g., heavy occlusion or high object density), the quality of candidate regions may degrade, reducing the effectiveness of the retrieval module.

Future work will focus on improving the efficiency and

robustness of candidate region generation, with the goal of producing high-quality masks under challenging conditions while reducing reliance on SAM. Additionally, further refinement of the <None> token routing mechanism could improve its ability to assess proposal quality, enabling more reliable decisions on when to use retrieved masks versus falling back to [SEG]-based prediction.

REFERENCES

[1] R. Hu, M. Rohrbach, and T. Darrell, "Segmentation from natural language expressions," in *European conference on computer vision*. Springer, 2016, pp. 108–124.

[2] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[3] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[4] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "Qwen2. 5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[5] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia, "Lisa: Reasoning segmentation via large language model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9579–9589.

[6] H. Rasheed, M. Maaz, S. Shaji, A. Shaker, S. Khan, H. Cholakkal, R. M. Anwer, E. Xing, M.-H. Yang, and F. S. Khan, "Glamm: Pixel grounding large multimodal model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 009–13 018.

[7] Z. Ren, Z. Huang, Y. Wei, Y. Zhao, D. Fu, J. Feng, and X. Jin, "Pixellm: Pixel reasoning with large multimodal model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 374–26 383.

[8] H. Yuan, X. Li, T. Zhang, Z. Huang, S. Xu, S. Ji, Y. Tong, L. Qi, J. Feng, and M.-H. Yang, "Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos," *arXiv preprint arXiv:2501.04001*, 2025.

[9] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[10] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[11] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in *International conference on machine learning*. PMLR, 2020, pp. 3929–3938.

[12] Z. Wei, W.-L. Chen, and Y. Meng, "Instructrag: Instructing retrieval-augmented generation via self-synthesized rationales," *arXiv preprint arXiv:2406.13629*, 2024.

[13] OpenAI, "ChatGPT," https://openai.com/blog/chatgpt/, 2023.

[14] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[15] S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie, "Eyes wide shut? exploring the visual shortcomings of multimodal llms," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9568–9578.

[16] Y. Li, Y. Zhang, C. Wang, Z. Zhong, Y. Chen, R. Chu, S. Liu, and J. Jia, "Mini-gemini: Mining the potential of multi-modality vision language models," *arXiv preprint arXiv:2403.18814*, 2024.

[17] S. Yang, Y. Chen, Z. Tian, C. Wang, J. Li, B. Yu, and J. Jia, "Visionzip: Longer is better but not necessary in vision language models," *arXiv preprint arXiv:2412.04467*, 2024.

[18] H. V. Team, P. Lyu, X. Wan, G. Li, S. Peng, W. Wang, L. Wu, H. Shen, Y. Zhou, C. Tang *et al.*, "Hunyuanocr technical report," *arXiv preprint arXiv:2511.19575*, 2025.

[19] C. Tang, Z. Han, H. Sun, S. Zhou, X. Zhang, X. Wei, Y. Yuan, J. Xu, and H. Sun, "Tspo: Temporal sampling policy optimization for long-form video language understanding," *arXiv preprint arXiv:2508.04369*, 2025.

[20] Z. Han, H. Sun, J. Xu, C. Tang, Y. Lei, X. Zhang, H. Sun, Z. He, and H. Sun, "Wat: Online video understanding needs watching before thinking," *arXiv preprint arXiv:2603.13412*, 2026.

[21] C. Tang, S. Zhou, H. Shi, and L. Wang, "Action hints: Semantic typicality and context uniqueness for generalizable skeleton-based video anomaly detection," *arXiv preprint arXiv:2509.11058*, 2025.

[22] H. Ding, S. Tang, S. He, C. Liu, Z. Wu, and Y.-G. Jiang, "Multimodal referring segmentation: A survey," *arXiv preprint arXiv:2508.00265*, 2025.

[23] Y.-C. Chen, W.-H. Li, C. Sun, Y.-C. F. Wang, and C.-S. Chen, "Sam4mllm: Enhance multi-modal large language model for referring expression segmentation," in *European Conference on Computer Vision*. Springer, 2024, pp. 323–340.

[24] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu *et al.*, "Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling," *arXiv preprint arXiv:2412.05271*, 2024.

[25] Z. Xia, D. Han, Y. Han, X. Pan, S. Song, and G. Huang, "Gsva: Generalized segmentation via multimodal large language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3858–3869.

[26] J. Wu, M. Zhong, S. Xing, Z. Lai, Z. Liu, Z. Chen, W. Wang, X. Zhu, L. Lu, T. Lu *et al.*, "Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks," *Advances in Neural Information Processing Systems*, vol. 37, pp. 69 925–69 975, 2024.

[27] D. Jang, Y. Cho, S. Lee, T. Kim, and D.-S. Kim, "Mmr: A large-scale benchmark dataset for multi-target and multi-granularity reasoning segmentation," *arXiv preprint arXiv:2503.13881*, 2025.

[28] S. Gong, Y. Zhuge, L. Zhang, Z. Yang, P. Zhang, and H. Lu, "The devil is in temporal token: High quality video reasoning segmentation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 29 183–29 192.

[29] H. Tang, C. Xie, H. Wang, X. Bao, T. Weng, P. Li, Y. Zheng, and L. Wang, "Ufo: A unified approach to fine-grained visual perception via open-ended language interface," *arXiv preprint arXiv:2503.01342*, 2025.

[30] M. Lan, C. Chen, Y. Zhou, J. Xu, Y. Ke, X. Wang, L. Feng, and W. Zhang, "Text4seg: Reimagining image segmentation as text generation," *arXiv preprint arXiv:2410.09855*, 2024.

[31] M. Zhu, Y. Tian, H. Chen, C. Zhou, Q. Guo, Y. Liu, M. Yang, and C. Shen, "Segagent: Exploring pixel understanding capabilities in mllms by imitating human annotator trajectories," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 3686–3696.

[32] J. Wang, Z. Wu, D. Huang, Y. Zheng, and H. Wang, "Unlocking the potential of mllms in referring expression segmentation via a light-weight mask decode," *arXiv preprint arXiv:2508.04107*, 2025.

[33] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "Referitgame: Referring to objects in photographs of natural scenes," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 787–798.

[34] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 11–20.

[35] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.