ROSE: Retrieval-Oriented Segmentation Enhancement

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

Figure 1. **Novel Emerging Segmentation.** (a) The cutoff date for Large Language Models' (LLMs) training data limits their knowledge of recent events. (b) Retrieval-Augmented Generation (RAG) enhances LLMs by retrieving up-to-date information from external databases. (c) Our task focuses on segmenting novel entities unrecognizable by existing models due to their absence in training data, and emerging entities that exist in the models' knowledge but require up-to-date external information.

Abstract

*Existing segmentation models based on multimodal large language models (MLLMs), such as LISA, often struggle with novel or emerging entities due to their inability to incorporate up-to-date knowledge. To address this challenge, we introduce the Novel Emerging Segmentation Task (NEST), which focuses on segmenting (i) novel entities that MLLMs fail to recognize due to their absence from training data, and (ii) emerging entities that exist within the model's knowledge but demand up-to-date external information for accurate recognition. To support the study of NEST, we construct a NEST benchmark using an automated pipeline that generates news-related data samples for comprehensive evaluation. Additionally, we propose **ROSE**: **R**etrieval-**O**riented **S**egmentation **E**nhancement, a plug-and-play framework designed to augment any MLLM-based segmentation model. ROSE comprises four key components. First, an Internet Retrieval-Augmented Generation module is introduced to employ user-provided multimodal inputs to retrieve real-time web information. Then, a Textual Prompt Enhancer enriches the model with up-to-date information and rich background knowledge, improving the model's perception ability for emerging entities. Furthermore, a Visual Prompt Enhancer is proposed to compensate for MLLMs'*

*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.

lack of exposure to novel entities by leveraging internet-sourced images. To maintain efficiency, a WebSense module is introduced to intelligently decide when to invoke retrieval mechanisms based on user input. Experimental results demonstrate that ROSE significantly boosts performance on the NEST benchmark, outperforming a strong Gemini-2.0 Flash-based retrieval baseline by 19.2% in gIoU.

1. Introduction

Segmentation models based on multimodal large language models (MLLMs), *e.g.*, LISA [19], SESAME [33], and READ [28], leverage MLLMs' [2, 4, 26, 31] reasoning abilities and world knowledge to address reasoning segmentation and achieve zero-shot capabilities. For example, when given an instruction "*Please segment the founder of SpaceX*", these models can identify the correct person in the image using MLLMs' knowledge. However, MLLMs require substantial computational resources for data collection, cleaning, and training, making frequent updates impractical. As a result, MLLM-based models struggle to incorporate newly emerging information. Since knowledge in the real world evolves rapidly, this limitation leads to failure cases in segmentation tasks involving novel or recently emerged entities. For example, LISA cannot accurately segment the current U.S. President due to its knowledge cutoff in 2023. To address this, it is promising to augment MLLM-based segmentation methods with Retrieval-Augmented Generation (RAG) techniques, as shown in Fig. 1 (b), enabling the model to effectively access and leverage up-to-date knowledge during inference.

In this work, we introduce a new segmentation task, **Novel Emerging Segmentation Task (NEST)**, to evaluate models' ability on segmenting novel and emerging entities, as shown in Fig. 1 (c). NEST requires generating binary segmentation masks based on user queries, with a focus on segmenting novel and emerging entities that are either unseen during training or require up-to-date knowledge for accurate interpretation. Novel entities refer to objects entirely absent from the MLLMs' training data. For example, models like LLaMA 3 [11], which have a knowledge cutoff at the end of 2023, are unable to recognize products such as the iPhone 17 pro max, which was released in 2025. Emerging entities, on the other hand, are included in the model's prior knowledge but evolve over time and require current context for accurate segmentation. For example, while LISA can segment Joe Biden and Donald Trump individually, it may fail to identify the current U.S. President due to outdated knowledge. Successfully addressing the NEST demands the following key capabilities: 1) retrieving up-to-date knowledge from the internet; 2) recognizing previously unseen entities; 3) applying the knowledge retrieved from the internet to generate accurate segmentation within visual inputs.

To support the study of novel emerging segmentation, we construct the **NEST benchmark**, containing over 1,500 image-question-answer-mask pairs. These samples are collected through an automated pipeline that continuously retrieves and updates the latest image-news pairs from the web. This automated pipeline provides a scalable and practical way for generating diverse and timely evaluation data tailored to novel and emerging entities. It enables continuous assessment of model performance as new concepts and objects emerge in the open world.

To tackle novel emerging segmentation, we propose **R**etrieval-**O**riented **S**egmentation **E**nhancement (**ROSE**), a plug-and-play method that can be integrated with any MLLM-based segmentation model (*e.g.*, LISA [19]). ROSE consists of four key components. First, the Internet Retrieval-Augmented Generation (IRAG) module enhances segmentation by retrieving real-time information from the web based on user-provided multimodal inputs. Second, the Textual Prompt Enhancer (TPE) supplements the model with precise target descriptions and rich contextual knowledge, improving segmentation performance on emerging objects. Third, the Visual Prompt Enhancer (VPE) further supports the recognition of novel entities by incorporating internet-sourced reference images. To ensure efficiency, the WebSense module adaptively determines whether retrieval is necessary based on the relevance of user inputs. By leveraging the latest online multimodal information, ROSE enables effective segmentation of both novel and emerging objects in a resource-aware manner.

This work makes the following key contributions:

- We introduce the Novel Emerging Segmentation Task (NEST), which challenges models to segment (i) novel entities unrecognizable by MLLMs and (ii) emerging entities requiring real-time information retrieval. This capability is essential for developing robust intelligent perception systems that can effectively adapt to and comprehend continuously evolving environments.
- We establish a ***NEST*** benchmark for novel emerging segmentation. Considering real-time data evolution, we develop an automated data engine that continuously constructs up-to-date datasets to evaluate models' novel emerging segmentation capabilities.
- We propose ***ROSE***, a plug-and-play method that augments any MLLM-based segmentation model with the ability to segment novel and emerging entities. ROSE integrates four components to retrieve and employ up-to-date multimodal information from the web.
- Experimental results shows that existing MLLM-based segmentation methods struggle to segment novel and emerging entities. Our proposed ROSE effectively addresses this limitation, outperforming a strong commercial retrieval baseline built on Gemini-2.0-Flash search by 19.2% in gIoU.

2. Related Work

Referring Expression Segmentation (RES). RES [5, 8, 9, 19, 28] aims to segment target objects in images based on textual descriptions. Early works [13, 22, 24] extract visual features using CNN and encode language expressions through LSTM. These extracted features are then fused through concatenation or other simple operations to create multi-modal representations. VLT [5, 6] first introduces the transformer architecture, which reformulates the RES task as an attention problem and proposes to use language features to query the vision features, generating results by decoding the transformer response. Liu *et al*. [25] proposes the Generalized Referring Expression Segmentation (GRES) task, supporting both the multi-target and empty-target scenarios. LLMs/MLLMs [1, 4, 11, 23, 26, 27] have revolutionized vision-language tasks by demonstrating remarkable capabilities in common-sense reasoning, opening up exciting new possibilities for RES [7, 8]. Building on these advances, LISA [19] innovatively introduces the [SEG] token, enabling the processing of expressions that require complex reasoning and common-sense knowledge. SESAME [33] employs model chaining and joint training to tackle false premise failures in MLLMs. READ [28] guides MLLMs on where to focus attention during interactive reasoning by treating similarity as reference points.

Retrieval-Augmented Generation (RAG). RAG [21, 36] has garnered significant attention in NLP [10, 12, 20] and multimodal communities [14, 21, 34]. Early methods like REALM [12] retrieve the top-k relevant snippets and use LLMs to generate k responses, combining them for QA tasks. Yu *et al*. [36] employ search engines to enhance LLMs on zero-shot knowledge-intensive tasks. Lazaridou *et al*. [20] apply few-shot prompting to leverage Google search results for factual and up-to-date QA. RA-CM3 [34] leverages external memory retrieval for text and image generation. SearchLVLMs [21] empowers MLLMs with real-time Internet search during inference.

3. NEST Dataset

The Novel Emerging Segmentation Task requires models to segment objects by acquiring up-to-date information. However, constructing a fixed, long-term dataset for evaluation is impractical, as such datasets may eventually be incorporated into the training of future MLLM-based segmentation models, introducing the risk of data leakage. Moreover, the collection, creation, and segmentation annotation of such datasets are labor-intensive, making it infeasible to maintain comprehensive, continually updated benchmarks for newly emerging entities. To address these challenges, we develop an automatic annotation pipeline that efficiently generates high-quality and continually up-to-date evaluation data for novel emerging segmentation task.

3.1. Automated NEST Data Engine

Each data sample in NEST consists of questions $\mathcal{Q}$, textual answers $\mathcal{A}$, reference images $\mathcal{I}_m$, and segmentation masks M, providing a comprehensive foundation for evaluating models' novel emerging segmentation capabilities.

Query Generation. As shown in Fig. 2, we construct the query collection Q primarily using Google Trends, supplemented with a small set of manually curated queries. Google Trends is a public platform that analyzes the popularity of search queries across regions and languages, surfacing trending keywords that reflect current user interests. While it offers valuable insights, the data is often noisy and skewed toward domains such as sports, entertainment, and politics. To improve domain diversity, we manually augment the query set with popular terms from technology, economics, and other underrepresented fields. However, many trending terms, *e.g*., abstract concepts like "Google stock", are not directly associated with segmentable objects. To address this, we employ large language models (LLMs) such as LLaMA 3 [11] and GPT-4o [15] to filter unqualified queries. This filtering process reveals that most viable segmentable queries correspond to people or products. The resulting filtered query set $\widetilde{Q}$ consists of concrete and clearly segmentable entities that maintain both topical relevance and category diversity, thereby enabling the effective construction of up-to-date segmentation samples.

Image Collection. We use the filtered query collection $\widetilde{Q}$ to retrieve corresponding images $\mathcal{I}$ from search engines for evaluating models' novel emerging segmentation capabilities. However, we observe that search engine results for many queries typically return images containing only a single visual entity, significantly reducing the task's complexity and failing to reflect realistic segmentation challenges. To address this limitation, we design an LLM-based query enhancement strategy that increases query complexity by encouraging the retrieval of images containing both the target entity and related entities. For example, instead of retrieving images featuring only "Mbappé", the enhanced query yields images that also include "Neymar" and "Messi", increasing visual ambiguity and segmentation difficulty. We submit both the original queries $\widetilde{Q}_o$ and the enhanced queries $\widetilde{Q}_e$ to search engines, resulting in two image sets: single-entity images $\mathcal{I}_s$ and multi-entity images $\mathcal{I}_m$. For $\mathcal{I}_s$, we apply CLIP-based feature clustering [29] and retain only the images from the largest cluster to ensure consistency and visual quality. These filtered images are used to support our auto-labeling pipeline. For $\mathcal{I}_m$, we apply an object detector [16] to filter out remaining single-entity samples. The resulting multi-entity images serve as the primary source for constructing the final segmentation dataset, ensuring higher task difficulty and diversity.

VQA Pair Construction. For each query retrieved from Google Trends, we also collect its associated metadata,

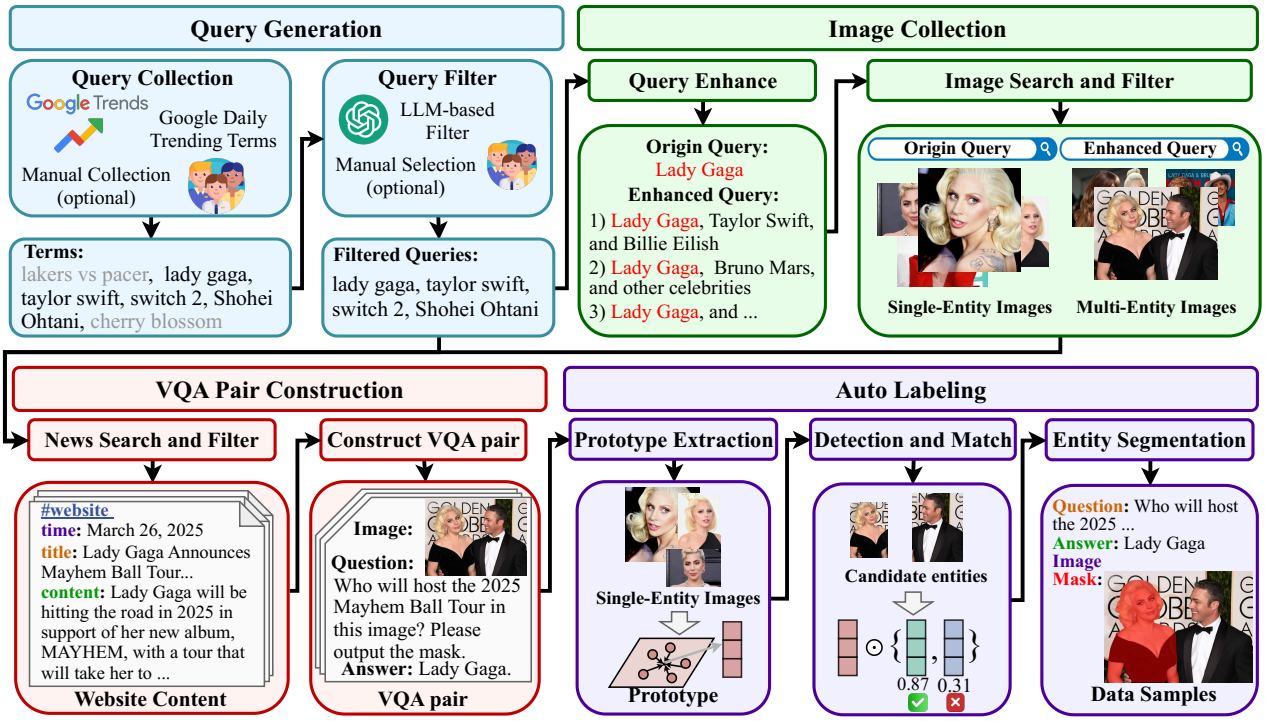

Figure 2. **NEST Data Engine.** We introduce an automated annotation pipeline that efficiently generates high-quality evaluation samples for the novel emerging segmentation task. The pipeline leverages time-specific queries to continuously collect news content and corresponding relevant images for constructing VQA pairs and automatically generating mask annotations, enabling a comprehensive and reliable evaluation of models' abilities to segment emerging entities.

including the query term, topic category, related keywords, and a set of related news articles. Each news item contains a link, publication timestamp, snippet, and other metadata. To reduce redundancy, we apply a temporal filtering strategy based on the observation that news articles published within a short time window often refer to the same event. Specifically, for articles published within a 3-day window, we retain only one representative article as the primary source for constructing visual question answering (VQA) pairs. We further refine this set by filtering out near-duplicate articles based on snippet similarity. To construct VQA samples, we employ LLMs to generate contextually relevant questions $\mathcal{Q}$ from the retained news content, where the original query $\widetilde{Q}_o$ serves as the answer $\mathcal{A}$. Our prompt design ensures that the generated questions are natural and do not explicitly mention the query term, thereby requiring genuine comprehension rather than simple keyword matching. Each resulting VQA sample is represented as a triplet $(\mathcal{Q}, \mathcal{A}, \mathcal{I}_m)$, where $\mathcal{Q}$ is the generated question, $\mathcal{A}$ is the answer, and $\mathcal{I}_m$ is the corresponding multi-entity image containing the target entity.

Auto Labeling. We employ an automated labeling pipeline to generate segmentation masks M for target entities, using the single-entity image set $\mathcal{I}_s$ associated with the ground-truth query $\widetilde{Q}_o$. First, we extract CLIP features f_s from the clustered single-entity images I_s to serve as the target entity representation. Next, we apply an object detector to the corresponding multi-entity images $\mathcal{I}_m$, identifying a set of entity proposals $\{E_i\}_{i=1}^n$. For each detected entity E_i, we crop its bounding region and extract CLIP features f_i. We then compute the cosine similarity between each f_i and the target representation f_s, selecting the entity with the highest similarity score. If the similarity exceeds a predefined threshold τ, the corresponding entity is selected as the target. We then pass its bounding box coordinates to the SAM mask decoder [18] to generate a high-quality segmentation mask M. This entire process is fully automated and requires no manual human intervention, greatly reducing the annotation burden while enabling scalable generation of up-to-date segmentation evaluation data.

3.2. NEST Dataset Analysis

We leverage our data engine to systematically collect and process web data by issuing queries to scrape news content between March 23, 2025 and April 11, 2025. The resulting *NEST* dataset contains 1,548 evaluation samples, primarily focusing on people and products across diverse domains including economics, technology, politics, entertainment, sports, and society. On average, each image contains 2.7 valid entities, increasing task complexity and mitigating hallucinations from large language models. Additionally, each image is paired with an average of 1.6 unique ques-

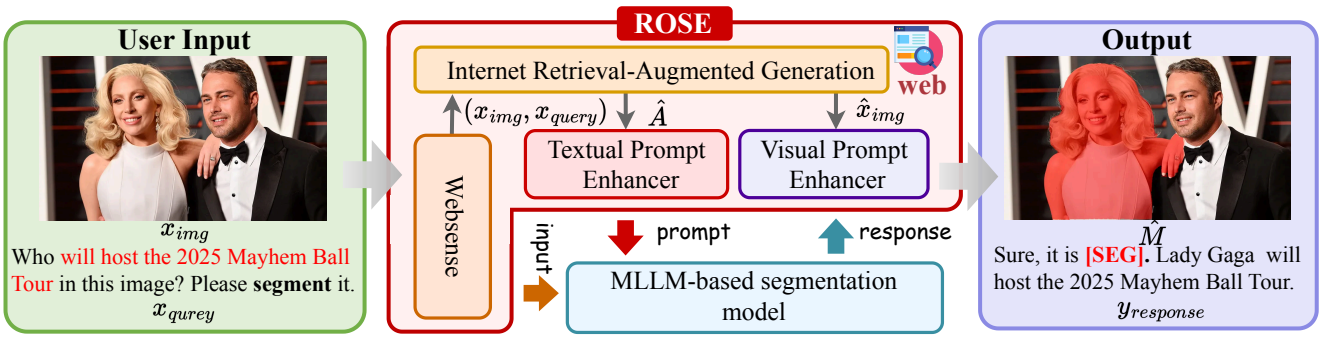

Figure 3. **Architecture overview of ROSE.** Given a user input (image and question), ROSE first employs the WebSense module to determine whether internet retrieval is needed. If so, the Internet Retrieval-Augmented Generation module retrieves relevant textual and visual data from the web. The retrieved content is then processed by the Textual and Visual Prompt Enhancer to generate enriched prompts for the MLLM-based segmentation model, which ultimately produces accurate segmentation masks for novel and emerging entities.

tions, enriching the diversity of query formulations. Further details are provided in Supplementary Materials.

4. Method

4.1. Architecture Overview.

The architecture of ROSE is shown in Fig. 3. Given an input image x_{img} and a user query x_{query}, ROSE first employs the WebSense module to determine whether internet retrieval is necessary. If needed, the IRAG module leverages x_{img} and x_{query} to retrieve answers $\hat{A}$ and relevant image $\hat{x}_{img}$ information from the internet. Then, the TPE module leverages $\hat{A}$ to generate an enhanced textual prompt $P = f(x_{query}, \hat{A}, K_{ext})$ to improve the MLLM-based segmentation model's ability to segment emerging entities, where K_{ext} represents extra background knowledge. For novel entities that may still be challenging to segment, the VPE module leverages the retrieved images $\hat{x}_{img}$ to extract their prototype features $\mathbf{f}_s$, which are then used to assist the model in segmenting novel entities. This comprehensive and unified method enables ROSE to effectively segment both novel entities that are entirely absent from MLLMs' training data and emerging entities that require continuously up-to-date information, thus expanding the segmentation capability from the original model's knowledge space K to an enhanced space $S = K \cup E$, where E represents the external knowledge database.

4.2. Internet Retrieval-Augmented Generation.

To retrieve up-to-date visual and textual information from the internet for enhanced segmentation, we develop an Internet Retrieval-Augmented Generation (IRAG) module based on LangChain [3] framework. This module processes user query x_{query} by first generating optimized search queries q using large language models (LLMs), then retrieving relevant content through search engines. The retrieved content is split into manageable chunks $\{C_i\}_{i=1}^{n}$, vectorized using embedding techniques $\mathbf{E}(C_i) \in \mathbb{R}^d$, and stored in a vector database $\mathcal{D} = \{(\mathbf{E}(C_i), C_i)\}_{i=1}^{n}$ for efficient retrieval. A map-reduce method with specialized prompts processes these chunks to extract the most relevant information, which is then synthesized into an answer candidate summary containing the potential answers $\{A_j\}_{j=1}^{m}$ to the user's query. However, since a question's answer is not always unique, the answer candidate summary may contain multiple valid answers. Therefore, we need to leverage the user-provided image x_{img} to narrow down the range of potential answers. To accomplish this, we employ Google Cloud Vision to analyze and extract multiple entities $\{E_k\}_{k=1}^{l}$ from the input image x_{img}. We choose not to use MLLMs for entity extraction because they lack the ability to accurately identify novel entities. Then, we compare the entities in the answer candidate summary $\{A_j\}_{j=1}^{m}$ with the detected entities $\{E_k\}_{k=1}^{l}$ to determine the correct answer $\hat{A}$. If no matching entities are found in the image, we select the entity with the highest confidence score from the answer candidate summary as the correct answer. Finally, we use $\hat{A}$ as a keyword to retrieve relevant images $\hat{x}_{img}$ from the internet.

4.3. Textual Prompt Enhancer.

To effectively leverage the retrieved answer $\hat{A}$ from IRAG, we design a textual prompt enhancer (TPE) to generate an enhanced textual prompt that improves the model's understanding of the target entity. This module integrates the query x_{query}, the answer $\hat{A}$ retrieved from the IRAG module, and the answer's extra background knowledge K_{ext} to create an optimized prompt. The background knowledge K_{ext} is obtained by retrieving the target's introduction from the Internet using $\hat{A}$ as the search term. By strategically combining the answer information from the IRAG module with comprehensive background knowledge, the module produces prompts that are both informative and directive, enabling the MLLM-based seg-

mentation model to accurately identify and segment the target object within the image.

4.4. Visual Prompt Enhancer.

VPE enhances the model's capability to segment novel entities by verifying and correcting results when the MLLM-based segmentation model fails to produce accurate segmentation. Using images $\hat{x}_{img}$ retrieved from the internet, we perform clustering and retain only the largest cluster. We then extract CLIP [29] features from these images to obtain a prototype feature representation $\mathbf{f}_s$. We compare this prototype feature with CLIP features extracted from the foreground region segmented by the MLLM-based segmentation model. A low similarity score indicates that the MLLM has failed to correctly identify the target entity. In such cases, we employ an object detector to identify multiple entities $\{E_i\}_{i=1}^{n}$ within the input image x_{img}. For each detected entity E_i, we extract CLIP features $\mathbf{f}_i$ from the cropped region. By computing the similarity between each $\mathbf{f}_i$ and the prototype feature $\mathbf{f}_s$, we identify the entity with the highest similarity score. When this score exceeds a predefined confidence threshold τ, we designate the corresponding entity as the target. Finally, we use the bounding box coordinates of the identified target as input to SAM's mask decoder [18], thereby generating a high-quality segmentation mask $\hat{M}$ for the target entity.

4.5. WebSense.

To reduce computational and network resources, the WebSense module is designed to intelligently determine whether the IRAG module should be activated based on x_{query}. While certain queries require real-time external knowledge retrieval to ensure accurate segmentation, others can be addressed using the internal knowledge of existing MLLMs without accessing external resources. This selective activation mechanism improves computational efficiency and reduces latency by invoking retrieval only when necessary. WebSense adopts a two-tier decision architecture. In the first stage, a lightweight rule-based filter quickly screens queries using predefined heuristics, *e.g.*, time-sensitive rules. For queries that are ambiguous or semantically complex, a large language model [11] is employed to perform deeper semantic analysis and classify whether retrieval is needed.

5. Experiments

Implementation Details. Our method is primarily built on the LangChain [3] framework, which enables efficient retrieval and reasoning capabilities. We employ Llama-3-8B [11] as our foundation language model, with a knowledge cutoff date of December 2023, which ensures no knowledge leakage for the NEST dataset. For prototype feature extraction, we employ CLIP-ViT-L/32 [29] to obtain rich visual representations. Additionally, we use YOLOv8 [16] as our object detector to identify potential regions of interest in the input images. All evaluations are conducted with a single NVIDIA 48G A6000 GPU.

Datasets. We evaluate the novel emerging segmentation capabilities of MLLM-based segmentation models using two datasets: NEST and NEST+. To enable more comprehensive and realistic evaluation, we construct NEST+ by combining NEST with ReasonSeg [19], RefCOCO, RefCOCO+[17], and RefCOCOg[35]. The composite dataset NEST+ supports joint evaluation across novel emerging segmentation, reasoning-based segmentation, and traditional referring segmentation tasks.

Baselines. We conduct experiments on several state-of-the-art MLLM-based segmentation models [19, 28, 33]. However, these models inherently lack internet retrieval functionality, limiting their ability to handle novel or time-sensitive queries. To ensure fair comparisons, we implement enhanced baselines by integrating commercial retrieval-augmented models, GPT-4o mini Search [15] and Gemini 2.0 Flash Search [31], which are equipped with built-in internet search capabilities.

Evaluation Metrics. Following [5, 19], we adopt two standard segmentation metrics: gIoU and cIoU. Additionally, we use Acc. to comprehensively evaluate the question-answering capability of RAG models.

5.1. Comparison with the State-of-the-Art Methods

Experiments on NEST Dataset. The novel emerging segmentation results on the NEST dataset are shown in Table 1. It is worth noting that existing works fail to handle this task, while our method successfully accomplishes it by retrieving up-to-date information, achieving more than 30% gIoU performance improvement. Unlike traditional referring segmentation, the novel emerging segmentation task presents a unique challenge by requiring models to identify and segment entities that have emerged after their training cutoff date. Only by leveraging the user's multimodal input to retrieve information from the internet can the model perform well on this task. Existing works have no effective way to identify entities beyond their knowledge cutoff, but our model explicitly exploits Retrieval-Augmented Generation (RAG) to more effectively address this challenge.

Furthermore, we also compare our method with vanilla two-stage baselines combining MLLM-based segmentation models with GPT-4o mini Search [15] and Gemini-2.0 Flash Search [31], which are advanced commercial LLMs with built-in internet retrieval capabilities. This baseline first uses them to generate an answer based on the input question, then employs MLLM-based segmentation models [19, 28, 33] to produce the segmentation mask. For the intermediate prompt to MLLM-based segmentation models, we use the template "*Please segment* {*answer*} *in this*

Table 1. **Novel Emerging Segmentation results on NEST dataset.** Furthermore, we partition NEST into a novel entity split and an emerging entity split based on LLaVA-v1.5-7B's [26] knowledge, which is the foundation model employed by LISA-7B [19], SESAME-7B [33], and READ-7B [28].

Method	RAG	Acc.	Novel Entity		Emerging Entity		Overall	
			gIoU	cIoU	gIoU	cIoU	gIoU	cIoU
CRIS [32]	✗	-	36.8	27.3	40.5	30.3	38.9	29.1
GRES [25]	✗	-	37.8	36.6	44.2	36.1	41.4	36.3
Grounded-SAM [30]	✗	-	39.0	31.7	53.8	40.3	47.4	36.7
SEEM [37]	✗	-	41.8	27.4	47.1	38.2	44.8	33.7
LISA-7B [19]	✗	-	38.4	28.5	56.5	47.5	48.7	39.3
SESAME-7B [33]	✗	-	11.1	7.9	14.5	14.0	13.1	11.6
READ-7B [28]	✗	-	19.2	17.6	25.1	22.0	22.5	20.2
LISA-7B+ GPT-4o mini Search [15]	✓	68.1	35.4	30.8	67.0	63.1	53.5	49.0
SESAME-7B+ GPT-4o mini Search	✓	67.8	27.8	22.6	42.3	41.7	36.1	33.8
READ-7B+ GPT-4o mini Search	✓	68.7	34.9	31.7	47.5	43.0	42.1	38.2
LISA-7B+ Gemini-2.0 Flash Search [31]	✓	69.6	35.2	29.6	67.8	65.3	53.8	49.3
SESAME-7B+ Gemini-2.0 Flash Search	✓	71.1	30.3	23.8	46.0	45.6	39.2	36.6
READ-7B+ Gemini-2.0 Flash Search	✓	70.0	37.1	31.6	50.2	44.9	44.6	39.3
LISA-7B + ROSE (ours)	✓	73.4	67.0	**65.7**	**77.5**	70.7	**73.0**	**68.6**
SESAME-7B + ROSE (ours)	✓	72.9	65.9	62.5	74.1	70.5	70.6	67.2
READ-7B + ROSE (ours)	✓	**74.2**	**67.1**	63.4	76.0	**71.7**	72.2	68.3

Table 2. **Mixed Dataset NEST+ Results.** ReasonSeg is sourced from [19], while RefSeg is derived from RefCOCO, RefCOCO+ [17] and RefCOCOg [35] datasets.

Method	RAG	NEST		ReasonSeg		RefSeg		Overall	
		gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU
CRIS [32]	✗	39.6	29.2	18.2	19.7	56.0	50.4	40.5	26.6
GRES [25]	✗	42.7	35.6	20.3	17.7	60.8	56.2	43.8	30.0
Grounded-SAM [30]	✗	47.8	34.4	20.6	15.6	42.4	31.4	43.8	27.7
SEEM [37]	✗	47.3	36.6	33.2	25.4	21.6	17.3	41.7	29.7
LISA-7B [19]	✗	51.1	39.0	42.5	44.2	54.9	53.7	50.9	40.6
SESAME-7B [33]	✗	14.1	11.2	30.9	28.7	62.7	61.7	25.5	17.8
READ-7B [28]	✗	22.3	18.3	49.7	54.2	63.9	59.9	33.4	28.3
LISA-7B + ROSE (ours)	✓	**75.3**	**67.4**	42.2	44.1	54.4	52.4	67.6	60.7
SESAME-7B + ROSE (ours)	✓	70.1	63.5	33.7	31.6	**68.4**	**66.3**	65.8	54.2
READ-7B + ROSE (ours)	✓	71.6	65.2	**50.3**	**54.9**	64.7	60.9	**67.9**	**62.4**

image." where *answer* is the response from the commercial LLM with built-in internet retrieval capabilities. As shown in Table 1, ROSE significantly outperforms this two-stage method for two main reasons: (1) Our TPE module provides more accurate target information and richer background knowledge, enhancing the MLLM's understanding of the target objects, and (2) our model leverages internet-sourced images to support novel entity segmentation , effectively addressing the novel entity segmentation problem. Additional experiments and detailed analysis are provided in Supplementary Materials.

Experiments on Mixed Dataset NEST+. To further evaluate the generalization ability of our method, we conduct experiments on a mixed dataset NEST+. The NEST+ simulates real-world scenarios involving real-time retrieval, reasoning, and traditional referring segmentation. As shown in Table 2, ROSE significantly enhances performance on the NEST split while maintaining competitive results on both the ReasonSeg and RefSeg splits. This demonstrates that ROSE improves the ability of novel emerging segmentation while maintaining performance on traditional tasks.

Table 3. **Ablation Study.** Impact of different ROSE components on novel emerging segmentation, conducted on the NEST dataset. The data split follows the same approach as in Tab. 1.

Method	Novel Entity		Emerging Entity		Overall	
	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU
LISA-7B [19]	38.4	28.5	56.5	47.5	48.7	39.3
+ROSE (IRAG only)	40.4	30.9	67.1	62.7	55.7	49.1
+ROSE (IRAG+TPE)	41.3	31.6	73.3	67.2	59.6	51.8
+ROSE (IRAG+VPE)	64.9	61.7	71.7	65.5	68.7	63.8
+ROSE (Full)	**68.6**	**67.2**	**79.4**	**72.3**	**74.7**	**70.1**

5.2. Ablation Study

Here, we perform an ablation study with LISA-7B [19] as the baseline. To ensure that each evaluation requires retrieval, the WebSense module is excluded from our ablation studies, allowing for a more comprehensive assessment of the model's ability to segment novel and emerging entities.

Effect of Internet Retrieval-Augmented Generation

Figure 4. **Qualitative results** comparing LISA [19], READ [28], and ROSE on novel and emerging entities. ROSE accurately segments unseen and newly emerging targets, while LISA struggles due to a lack of up-to-date knowledge or inability to recognize new entities.

Module (IRAG). As shown in Table 3, adding the IRAG module brings effective improvements over the baseline method LISA-7B, with overall gIoU increasing by 7.0% and cIoU increasing. This substantial enhancement confirms that retrieving up-to-date information from the internet is crucial for novel emerging segmentation, as the model cannot rely solely on pre-trained knowledge to identify and segment novel or emerging entities.

Effect of Textual Prompt Enhancer (TPE). As shown in Table 3, comparing IRAG-only with IRAG+TPE, we observe that TPE improves performance on emerging entities (gIoU +6.2% and cIoU +4.5%). This demonstrates that well-structured textual prompts effectively integrate retrieved knowledge with the original instruction.

Effect of Visual Prompt Enhancer (VPE). As shown in Table 3, comparing IRAG-only with IRAG+VPE, we observe that VPE significantly improves performance on novel entities (cIoU +24.5%) and overall performance (gIoU +13.0%). This indicates that enhancing the model's visual understanding through retrieved images significantly improves its ability to segment novel entities by providing visual references that complement the textual information.

5.3. Qualitative Results

Fig. 4 presents qualitative results demonstrating the effectiveness of ROSE. We present four examples comparing ROSE with LISA [19] and READ [28] on two key challenges in novel emerging segmentation: novel entities and emerging entities. For novel entities in the first two rows, ROSE accurately segments the referred targets, while LISA fails due to its inability to recognize unseen entities. For example, in the 2nd row, LISA produces no output for the novel entity Xiaomi SU7, whereas ROSE correctly segments it. For emerging entities in the last two rows, LISA similarly struggles due to outdated knowledge. In the 4th row, given the question "*Which MLB player hit a clutch three-run homer for the Dodgers on May 9, 2025?* ", LISA segments the wrong person, while ROSE correctly identifies and segments the actual player. More visualization results are provided in the Supplementary Materials.

6. Conclusion

We introduce the novel emerging segmentation task (NEST), which requires segmenting (i) novel entities absent from MLLMs' training data and (ii) emerging entities that demand up-to-date external knowledge. To support this, we construct the NEST dataset using an automated pipeline that collects real-time image–news pairs. We propose ROSE, a plug-and-play approach that enhances MLLM-based segmentation models through real-time internet retrieval. Extensive experiments show that ROSE significantly improves performance on novel emerging segmentation while maintaining competitive results on standard benchmarks. Our work highlights the potential of combining retrieval with multimodal models for real-time understanding of newly emerging entities.

Acknowledgement. This work was supported by the Science and Technology Commission of Shanghai Municipality (No. 25511103600) and the National Natural Science Foundation of China (NSFC) under Grant No. 62472104.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[2] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023. 2

[3] Harrison Chase. Langchain: Building applications with large language models. 2022. 5, 6

[4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024. 2, 3

[5] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *ICCV*, 2021. 3, 6

[6] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. VLT: Vision-language transformer and query generation for referring segmentation. *IEEE TPAMI*, 45(6), 2023. 3

[7] 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. 3

[8] 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. 3

[9] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Yu-Gang Jiang. GREx: Generalized referring expression segmentation, comprehension, and generation. *IJCV*, 2026. 3

[10] Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024. 3

[11] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2, 3, 6

[12] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *ICML*, 2020. 3

[13] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *ECCV*, 2016. 3

[14] Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *CVPR*, 2023. 3

[15] 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. 3, 6, 7

[16] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023. 3, 6

[17] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 6, 7

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 4, 6

[19] 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. 2, 3, 6, 7, 8

[20] Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. Internet-augmented language models through few-shot prompting for open-domain question answering. *arXiv preprint arXiv:2203.05115*, 2022. 3

[21] Chuanhao Li, Zhen Li, Chenchen Jing, Shuo Liu, Wenqi Shao, Yuwei Wu, Ping Luo, Yu Qiao, and Kaipeng Zhang. Searchlvlms: A plug-and-play framework for augmenting large vision-language models by searching up-to-date internet knowledge. In *NeurIPS*, 2024. 3

[22] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *CVPR*, 2018. 3

[23] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 3

[24] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *ICCV*, 2017. 3

[25] Chang Liu, Henghui Ding, and Xudong Jiang. GRES: generalized referring expression segmentation. In *CVPR*, 2023. 3, 7

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2024. 2, 3, 7

[27] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 3

[28] Rui Qian, Xin Yin, and Dejing Dou. Reasoning to attend: Try to understand how <SEG> token works. In *CVPR*, 2025. 2, 3, 6, 7, 8

[29] 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. 3, 6

[30] 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. 7

[31] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2, 6, 7

[32] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *CVPR*, 2022. 7

[33] 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. 2, 3, 6, 7

[34] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Retrieval-augmented multimodal language modeling. In *ICML*, 2022. 3

[35] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016. 6, 7

[36] Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. Generate rather than retrieve: Large language models are strong context generators. In *ICLR*, 2022. 3

[37] 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. 7