URECA 💡: Unique Region Caption Anything

Sangbeom Lim[1*] Junwan Kim[2*] Heeji Yoon[3] Jaewoo Jung[3] Seungryong Kim[3†]

[1]Korea University [2]Yonsei University [3]KAIST AI

https://cvlab-kaist.github.io/URECA

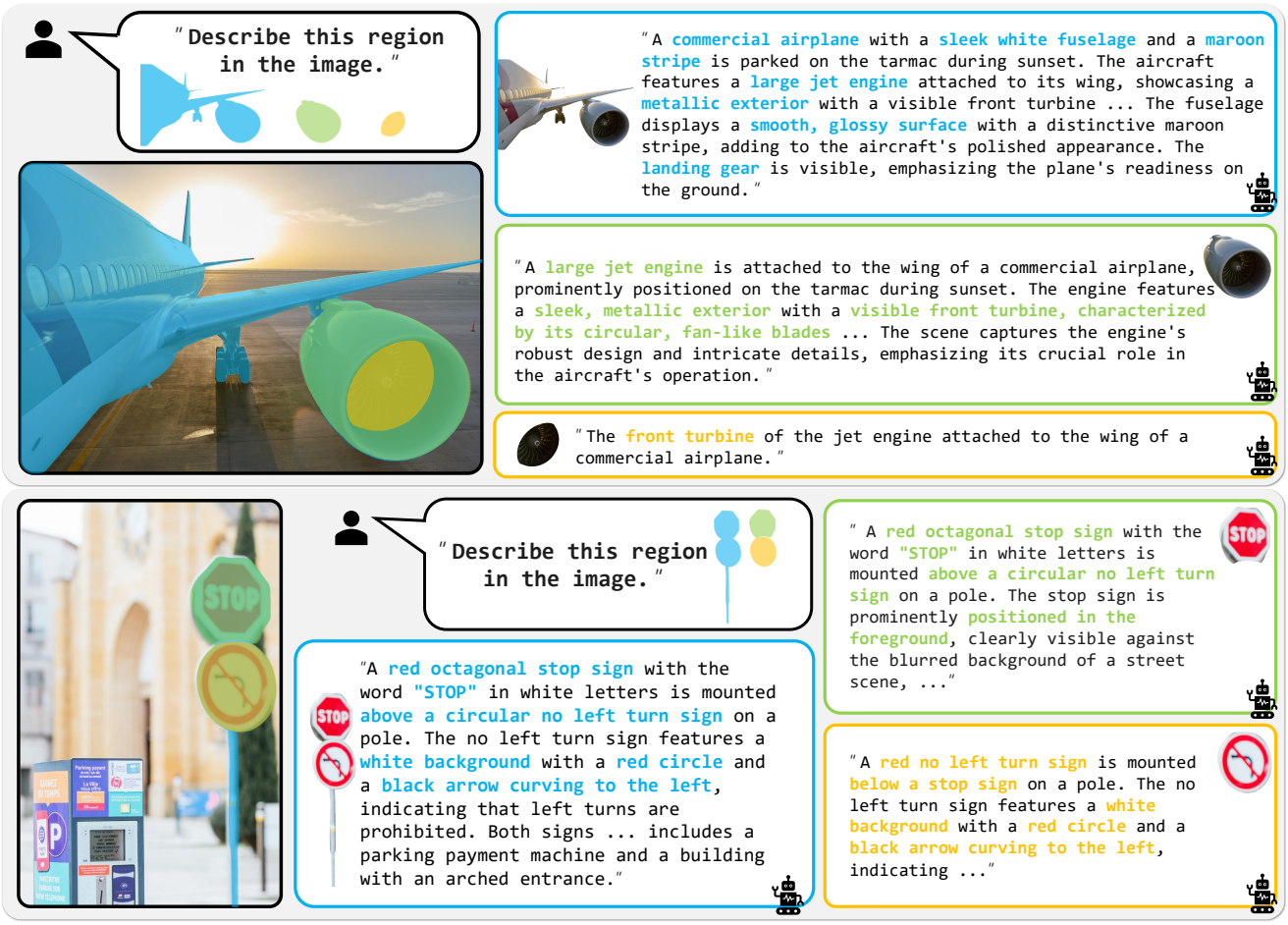

Figure 1. **Unique Region Caption Anything.** We introduce URECA dataset, a novel region-level captioning dataset designed to ensure caption uniqueness and support multi-granularity regions. Each caption in our benchmark is uniquely mapped to its corresponding region, capturing distinctive attributes that differentiate it from surrounding areas. Moreover, we show that our proposed model trained on our dataset effectively generates unique captions for regions at any level of granularity.

Abstract

Region-level captioning aims to generate natural language descriptions for specific image regions while highlighting their distinguishing features. However, existing methods struggle to produce unique captions across multi-granularity, limiting their real-world applicability. To address the need for detailed region-level understanding, we introduce URECA dataset, a large-scale dataset tailored for multi-granularity region captioning. Unlike prior datasets

*These authors contributed equally.
†Corresponding author.

that focus primarily on salient objects, URECA dataset ensures a unique and consistent mapping between regions and captions by incorporating a diverse set of objects, parts, and background elements. Central to this is a stage-wise data curation pipeline, where each stage incrementally refines region selection and caption generation. By leveraging Multimodal Large Language Models (MLLMs) at each stage, our pipeline produces distinctive and contextually grounded captions with improved accuracy and semantic diversity. Building upon this dataset, we present URECA, a novel captioning model designed to effectively encode multi-granularity regions. URECA maintains essential spatial properties such as position and shape through simple yet impactful modifications to existing MLLMs, enabling fine-grained and semantically rich region descriptions. Our approach introduces dynamic mask modeling and a high-resolution mask encoder to enhance caption uniqueness. Experiments show that URECA achieves state-of-the-art performance on URECA dataset and generalizes well to existing region-level captioning benchmarks.

1. Introduction

Region-level captioning aims to describe a specific region in natural language while highlighting its distinguishing features compared to other regions. Although previous approaches have shown strong performance in describing target regions, they have struggled to generate *distinguishable captions across multiple granularities*. For real-world applications, it is crucial to describe regions at any level of granularity.

Despite its importance, generating unique captions across multiple levels of granularity remains underexplored. The primary challenges in multi-granularity captioning include accurately localizing the user-specified region and capturing its unique attributes (e.g., color, relative position, and shape) to clearly distinguish it from surrounding areas [30, 47, 50]. Furthermore, generating truly unique regional captions [35, 59] is particularly challenging, as most existing methods [19, 20, 36, 39, 57, 61, 65, 66] primarily describe target regions without sufficiently capturing their distinctive attributes regarding their contextual surroundings. Additionally, even though several studies [7, 28, 29, 42] have explored regional captioning, we find that they struggle to produce truly unique and context-aware captions due to limitations in both model design and dataset availability.

Regarding model design, previous approaches [9, 31, 51, 63] have explored generating captions on salient regions by providing region coordinates through natural language descriptions [5, 29], marking regions directly on the image (e.g., mask contours, ellipses, bounding boxes, triangles, scribbles, points, arrows, and masks) [4, 41, 55], or utilizing visual RoI features [12, 16]. Although these methods have demonstrated notable performance, they come with several drawbacks on solving unique multi-granularity region captioning: they may alter the original image, make it difficult to distinguish between colored contours, and fail to capture global relationships within the image [43], ultimately hindering the model's ability to generate unique captions.

On the other hand, we show that the naïve use of existing captioning datasets [6, 14, 15, 18, 22, 31, 38, 39, 45, 49, 56, 61] is also incompatible with unique multi-granularity captioning. Existing datasets primarily focus on salient regions [39, 45, 49, 56] or rely on bounding box annotations [14, 31], which limits the model's ability to describe less salient properties. Additionally, most captions in these datasets tend to be overly general [22, 38, 61] (e.g., "a person standing"), leading to duplicate annotations for multiple regions within the same image. This lack of specificity prevents models from distinguishing similar regions effectively, ultimately hindering their ability to generate faithful descriptions of user-defined regions.

As current datasets have not explored on both generating unique caption and considering multi-granularity, our paper first presents a novel data curation pipeline on generating unique captions on multi-granularity regions. To generate unique captions for regions while considering their hierarchical relationships, our automated data curation pipeline annotates regions in a stage-wise manner using a *mask tree* structure. The mask tree captures hierarchical dependencies by representing subset and superset relationships among regions, as shown in Figure 2. By leveraging this structure, our pipeline efficiently generates unique captions for target regions based on their corresponding tree nodes.

Utilizing our data pipeline, we propose **URECA dataset**, a dataset tailored for unique captioning on multi-granularity regions. Unlike previous captioning datasets [6, 14, 18, 22, 31, 38, 39, 45, 49, 56, 61] which are solely limited to salient regions within an image and contain simple descriptions, our dataset covers a broader range of objects, parts, and backgrounds each with a unique caption. Each caption in our dataset is uniquely mapped to a single region, ensuring a unique correspondence between regions and captions across various granularities.

Based on our dataset, we then propose a novel model architecture **URECA**, which effectively conditions regions of interest for the captioning model without losing the region's details. To consider multi-granularity regions, it is essential to encode various size of regions, which previous methods often struggled with [4, 12, 18, 60, 65]. To achieve this, we introduce a mask encoder network and dynamic mask modeling approach that extracts mask features while preserving essential properties such as position and shape even on the multi-granularity regions. To demonstrate the effectiveness of our proposed **URECA dataset** and **URECA** in generating unique multi-granularity region captions, we

Dataset	Simple caption	Dense caption	Region caption	Multi-granularity	Unique caption
RefCOCOg [56]	✓	✗	✓	✗	✗
Visual Genome [22]	✓	✗	✓	✗	✗
PACO [38]	✓	✗	✓	✗	✗
Partimagenet [6]	✓	✗	✓	✗	✗
PRIMA [45]	✓	✓	✗	✗	✗
LLaVA-115K [29]	✓	✓	✗	✗	✗
Arcana [42]	✓	✓	✓	✗	✗
Osprey [61]	✓	✓	✓	✗	✗
I Dream My Painting [10]	✓	✓	✓	✗	✗
GRIT [36]	✓	✓	✓	✗	✗
LiSA [23]	✓	✓	✓	✗	✗
USE [49]	✓	✓	✗	✓	✗
SegCAP [67]	✓	✓	✓	✓	✗
GranD [39]	✓	✓	✓	✓	✗
URECA dataset (Ours)	✓	✓	✓	✓	✓

Table 1. **Statistical comparison of previous captioning datasets and URECA dataset in region-level captioning.** The comparison covers different types of captions, including simple captions (*e.g.*, [15, 38]), dense captions (*e.g.*, [29, 45]), region captions (*e.g.*, [10, 22, 23, 42, 51, 56, 61]), and multi-granularity captions (*e.g.*, [39, 49, 67]). While these datasets provide varying levels of detail, URECA dataset is the only dataset that offers distinctive dense captions and handles multi-granularity regions effectively.

further validate our model on a test set that has undergone an additional quality assurance stage to ensure the quality of the annotated captions produced by our automated pipeline. Notably, URECA achieves state-of-the-art performance on our URECA dataset test set while also demonstrating strong performance on traditional region-level captioning datasets such as Visual Genome [22] and RefCOCOg [56] datasets. We also show that further fine-tuning existing captioning models on URECA dataset enables better multi-granularity captioning.

In summary, we make the following contributions:

- We introduce a large-scale multi-granularity captioning dataset that ensures unique region-caption mapping by covering a diverse range of objects, parts, and backgrounds beyond salient regions.
- We present a novel captioning model designed to handle multi-granularity regions through dynamic mask modeling and a high-resolution mask encoder that preserves essential region properties.
- Extensive experiments demonstrate that our model achieves state-of-the-art performance on our test dataset and shows strong generalization on benchmark datasets, including Visual Genome [22] and RefCOCOg [56].

2. Related Work

Multi-modal large language model. Large Language Models (LLMs) have demonstrated pioneering performance in instruction following capabilities, integrating diverse knowledge from extensive datasets, and performing complex reasoning tasks. However, a significant limitation of LLMs is their reliance solely on natural language inputs. To address this, LLaVA [29] was the first to explore the integration of image and text modalities by representing visual features as visual tokens. Building upon this, models such as Flamingo [1] and BLIP-2 [26] have further advanced Multimodal Large Language Models (MLLMs) by incorporating powerful visual backbones. These models effectively bridge the two modalities and have shown strong performance in tasks like image captioning and visual question answering. Building on these advancements, recent efforts have aimed to extend these models to handle more complex tasks, including reasoning over segmentation [23, 40], optical character recognition [8, 48], and grounding [13, 37, 39, 49, 67].

Region-level vision language model. Although MLLMs have demonstrated impressive image understanding capabilities, generating captions for specified regions remains a challenging task. LLaVA [29] and MiniGPT-2 [5] have explored conditioning given regions by translating bounding box coordinates into natural language. However, these models heavily rely on the MLLMs' ability to interpret bounding box coordinates accurately. Other approaches [4, 41, 55] have attempted to overlay regions directly onto the image. While this method is straightforward to implement, it alters the original image, making it difficult for MLLMs to reference the unmodified content. To address this issue without modifying the original image, some methods have explored directly modeling the coordinates of the regions or feature pooling conditioned on bounding boxes [9, 31, 51, 63]. Although pooling features from the

bounding box has improved performance, these approaches often struggle to accurately capture user intent, particularly when objects overlap. Mask-based feature pooling [12, 16] provides more precise localization information by avoiding ambiguous bounding box indications. However, it is typically performed on low-resolution image features and excessively aggregates information, leading to the loss of fine-grained details such as shape and boundaries. In extreme cases, small-region masks in high-resolution images may disappear entirely during this process, resulting in the loss of meaningful features.

Moreover, none of the prior works have effectively addressed the challenge of generating captions that precisely localize user-intended regions while capturing their unique attributes at any granularity. This is primarily due to the lack of a suitable dataset and the absence of architectures designed for this task. To bridge this gap, we propose an automated data generation pipeline that ensures the inclusion of unique captions while considering multi-granularity regions. Additionally, our architecture effectively handles such multi-granularity regions, preserving their original attributes and capturing global relationships among regions.

3. URECA Dataset

Previous research has made significant progress in generating dense region captions; however, approaches focusing on multi-granularity regions remain scarce. When considering the granularity of regions, distinguishing their unique attributes becomes crucial [30, 35, 47, 50], as visually similar regions frequently appear within an image. Existing approaches have struggled to generate truly unique captions for regions, often producing generic descriptions despite clear visual differences.

This tendency to generate generic captions contradicts human perception, as humans naturally recognize and describe regions based on distinctive attributes like color, position, and shape. However, existing captioning datasets often lack such specificity, and training models on such generic captions that do not emphasize regional uniqueness can contribute to the *mode collapse* problem [50], where models fail to generate diverse and informative captions.

To address this lack of specificity in existing datasets, we propose URECA dataset, a dataset designed to enhance models' ability to generate unique captions for given multi-granularity regions. Our dataset is generated through an automated data pipeline that creates and verifies captions in a stage-wise manner. Specifically, we build our dataset using the publicly available SA-1B dataset, which offers high-resolution images and multi-granularity regions. To further ensure caption quality in the test set, we incorporate a verification step using GPT-4o [32] as part of the pipeline.

Data annotation pipeline. To generate unique captions that effectively capture multi-granularity, it is crucial to consider both target and non-target regions. Captions that focus solely on the target region often become overly localized and repetitive, making it difficult to distinguish between similar regions. To address this, we structure hierarchical relationships between regions, ensuring that captions incorporate broader contextual information.

At the core of our approach is a mask tree, constructed based on Intersection-over-Union (IoU). This hierarchical structure organizes regions into subset-superset relationships, allowing us to systematically capture dependencies between different regions. This hierarchical structure enables a comprehensive understanding of region dependencies at both global and local levels, ensuring the generation of unique captions.

This process follows a structured sequence of four stages, as illustrated in Figure 2:

1. **Mask tree generation.** We first construct a mask tree to represent the hierarchical relationships among masks in an image. By comparing the IoU between masks, we can determine their relationships (i.e., superset or subset) within the hierarchy.
2. **Top-down generation.** To ensure that contextual information is effectively incorporated into each node's caption, we generate captions in a top-down manner. In this process, each node refers to its parent node to maintain hierarchical consistency. Specifically, we generate short captions using our annotation MLLM, InternVL2.5-38B [7], for each node by referring captions from the parent node and two types of images that represent the target region: a cropped image of the target region with non-target areas blurred based on the mask [54], and a cropped image of the parent region, where the target region is contoured while non-target areas within the parent region are blurred.
3. **Bottom-up generation.** To ensure that parent nodes have unique captions incorporating relevant details from their child nodes while maintaining contextual coherence, we generate captions in a bottom-up manner. In this process, the parent node refers to its children's captions to generate a more informative and unique caption. Specifically, we aggregate the captions of all child nodes and use our annotation MLLM to generate a refined caption based on the aggregated captions, the parent node's short caption, and an image where the target region is contoured within the full image to preserve its spatial context.
4. **Uniqueness refinement.** To further ensure visually similar regions have distinguishable captions, we introduce a uniqueness refinement process based on image feature similarity using DINOv2 [33]. In this stage, similar-looking regions are identified using image features and

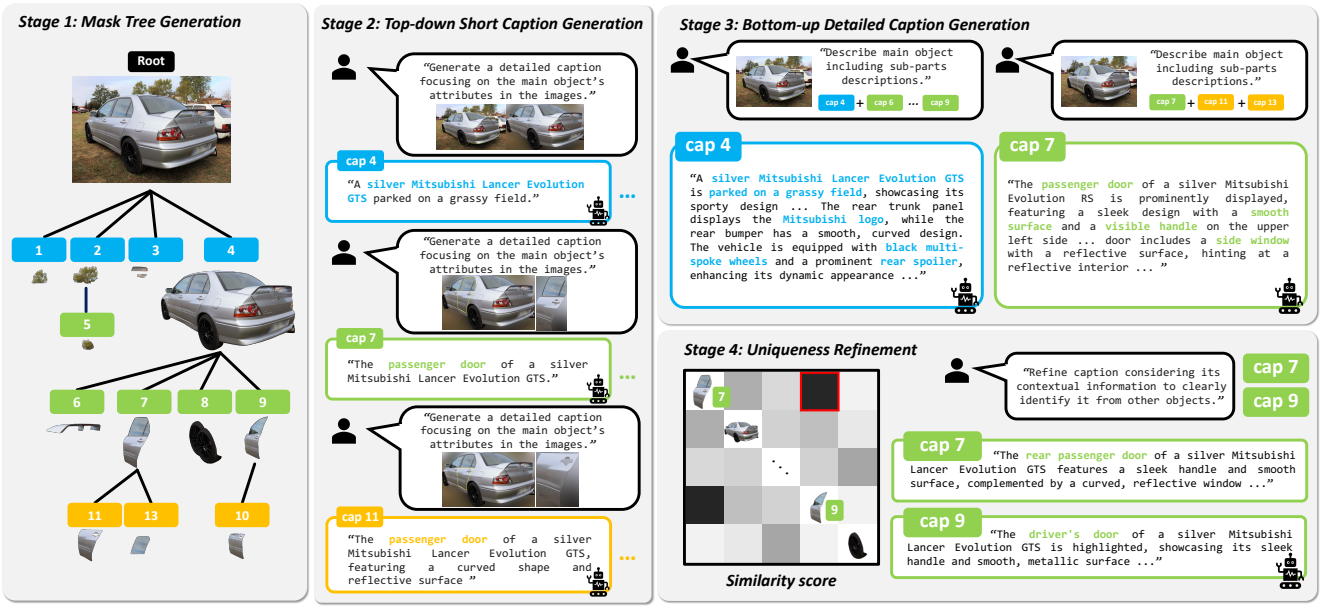

Figure 2. **Automated data curation pipeline of URECA dataset.** Our pipeline consists of four key stages to generate unique captions for multi-granularity regions. In Stage 1, we construct a mask tree that captures hierarchical relationships between regions. Stage 2 generates short captions based on the parent node. Stage 3 aggregates captions from child nodes, and Stage 4 ensures that each node is assigned a unique caption. Best viewed in zoomed-in.

marked in the image with contours and indexed bounding boxes [53]. Our annotation MLLM then generates a unique caption by explicitly differentiating the target region from other visually similar regions.

Data statistics. We conducted a statistical comparison between previous captioning datasets and URECA dataset. Table 1 highlights their capabilities in region-level captioning. Simple caption refers to datasets [15, 38] that provide basic descriptions, often incorporating object classes in the captions. Dense caption represents datasets [29, 45] that include multiple attributes, offering more detailed descriptions of the region. Additionally, datasets [10, 22, 23, 42, 51, 56, 61] where captions are explicitly aligned with specific regions fall under the region caption category. As multi-granularity captioning becomes increasingly relevant for real-world applications, recent datasets [39, 49, 67] have started to incorporate this aspect. However, none of the existing datasets fully capture all these aspects with captions that describe distinctive attributes of the region while maintaining multi-granularity. Among them, URECA dataset stands out as a unique dataset providing distinct dense captions while effectively handling multi-granularity regions.

Evaluation set. To ensure the quality of the test dataset when evaluating unique captioning on multi-granularity regions, we additionally implemented a verification stage during the test set generation process. As state-of-the-art MLLMs have demonstrated performance comparable to human annotators' preferences [11, 24, 52], we utilized GPT[1], which is widely adopted to simulate human annotators for data generation tasks. Further details about the dataset pipeline can be found in Appendix.

[1] gpt-4o-mini-2024-07-18

4. URECA

The overall architecture of URECA is illustrated in Figure 3. It builds upon the LLaVA [29] architecture, which treats visual features as tokens. Inspired by this approach, we adopt a similar strategy by representing mask features as tokens. To achieve this, we integrate a mask encoder that localizes the target region while preserving essential mask details such as size, position, and shape.

4.1. Mask Encoder

To capture the distinctive features of a target region, it is essential to leverage both local information and the global context of the image. A representation of the region should function as a localizer rather than a constraint on the region. Additionally, the representation of the region should encode information such as size, position, and shape without ambiguity or loss. Masks inherently provide this capability, whereas previous methods such as contours, bounding boxes, and text coordinates do not. Therefore, encoding mask information without loss is crucial for generating dense and distinctive captions.

To achieve this, we introduce a mask encoder that exclusively encodes the mask without altering the original image, thereby preserving the mask's unique attributes. Our mask

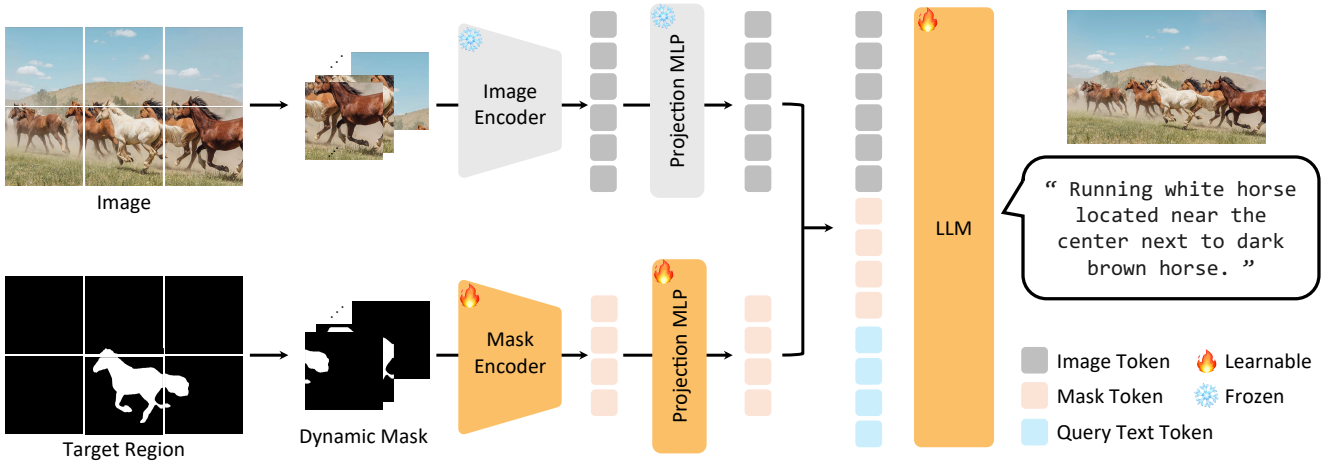

Figure 3. **URECA architecture.** URECA enables users to generate unique captions that describe distinctive attributes of any region. The mask encoder effectively encodes multi-granularity regions while preserving their identity. The mask token serves as a localizer, guiding the LLM to generate region-specific captions based on the image and query token.

encoder transforms the binary mask into a sequence of tokens through multiple convolutional layers, resulting in a set of mask tokens. These mask tokens are then integrated with image tokens within the MLLM, enabling the mask to function effectively as a localizer while maintaining precise region-specific information. More precisely, our mask encoder performs as:

$$\mathbf{F} = \phi(\mathbf{M}) \in \mathbb{R}^{N \times D}, \tag{1}$$

where $\mathbf{M} \in \{0,1\}^{H \times W}$ denote the input binary segmentation mask, and H and W represent the height and width of the mask, respectively. The encoding process is performed by a mask encoder, represented by the function $\phi(\cdot)$. Our mask encoder maps the binary mask $\mathbf{M}$ to a feature representation $\mathbf{F} \in \mathbb{R}^{N \times D}$, where N denotes the number of spatial tokens, and D is the feature embedding dimension.

Unlike traditional mask feature pooling, our mask encoder tokenizes the mask without aggregation, preserving fine-grained spatial details and capturing multi-granularity regions. The resulting mask tokens contain both local and global details of the mask, enabling our model to accurately locate regions and generate clear, unique captions.

4.2. Dynamic Mask Modeling

Naïvely using an encoder that receives fixed-size inputs requires mask resizing, leading to the loss of fine-grained region details. However, preserving these details is particularly important, as multi-granularity captioning relies on masks of diverse scales. Therefore, generating mask tokens directly from high-resolution inputs is essential.

Thanks to the success of MLLMs that accept dynamic-length inputs, we leverage this capability in our mask modeling. To provide a precise view during mask modeling and alleviate the challenges caused by extensive resizing, we propose dynamic masking which split the original high-resolution mask into multiple sub-masks. This approach allows the length of mask tokens to be dynamically adjusted based on resolution, ensuring a more accurate and flexible representation.

Specifically, before passing the mask through the encoder, we first apply a dynamic masking process to split the original mask $\mathbf{M} \in \{0,1\}^{H \times W}$ into multiple sub-masks $\mathbf{M}_{\text{split}}$ as follows:

$$\mathbf{M}_{\text{split}} = \text{Split}(\mathbf{M}) \in \{0,1\}^{N_s \times H' \times W'}. \tag{2}$$

These sub-masks are obtained by dividing the original mask into smaller regions, each having a size of $H' \times W'$. The number of sub-masks N_s depends on the pre-defined splitting strategy. This process ensures that the mask encoder receives high-resolution and localizes information while preserving the global context of the original mask. The sub-masks $\mathbf{M}_{\text{split}}$ are then passed through the mask encoder, resulting in the tokenized feature representation $\mathbf{F}_{\text{split}} \in \mathbb{R}^{N_s \times D}$, where D is the feature embedding dimension. This dynamic masking step allows for finer localization of regions and helps capture multi-granularity features before encoding them into a compact representation.

5. Experiments

5.1. Quantitative Results

We report the performance of URECA on URECA dataset as well as previous benchmark datasets [22, 56]. All results are evaluated using an 8B language model trained exclusively on the URECA dataset.

Unique multi-granularity region captioning. In Table 2, we present the performance comparison on URECA

Models	BLEU@1	BLEU@2	BLEU@3	BLEU@4	ROUGE	METEOR	BERTScore
None	17.06	7.63	3.14	1.20	17.86	27.72	62.68
Contour	17.10	7.13	2.63	1.01	19.95	25.49	63.29
Crop	18.43	7.53	2.45	0.85	19.73	26.45	63.63
SCA [19]	22.76	13.58	6.97	3.88	30.76	24.87	70.67
KOSMOS-2 [36]	30.31	18.12	9.96	5.55	34.19	32.94	72.64
OMG-LLaVA [65]	34.01	21.88	13.51	8.46	38.14	37.29	74.68
ViP-LLaVA [4]	34.17	22.07	13.96	9.00	38.17	37.68	74.62
URECA (Ours)	**36.56**	**23.84**	**15.42**	**9.98**	**38.95**	**41.25**	**75.11**

Table 2. **Performance comparison of URECA with baseline methods and previous models on various evaluation metrics, including BLEU [34], ROUGE [27], METEOR [3], and BERTScore [64].** The results show that URECA outperforms other methods across all metrics on URECA testset, demonstrating its superior ability to generate unique captions for multi-granularity regions. Note that comparison methods are all trained on URECA dataset.

Models	RefCOCOg	Visual Genome
ControlMLLM [46]	14.0	-
Kosmos-2 [36]	14.1	-
GRiT [51]	15.2	17.1
SLR [58]	15.9	-
GLaMM [39]	15.7	17.0
OMG-LLaVA [65]	15.3	-
ViP-LLaVA [4]	16.6	-
Groma [31]	16.8	16.8
RegionGPT [12]	16.9	17.0
Omni-RGPT [16]	**17.0**	17.0
URECA (Zero-Shot)	16.1	**18.4**

Table 3. **Quantitative results on region-level captioning task.** Performance comparison on the METEOR [3] for the RefCOCOg [56] and Visual Genome [22] datasets. (Zero-Shot) refers to zero-shot transfer.

dataset, a dataset specifically designed to evaluate unique multi-granularity region captions, alongside previous methods. To demonstrate the effectiveness of our approach, we implemented a baseline by running a naïve MLLM [7] on URECA dataset. "None" refers to providing the MLLM with only the image, without any explicit region marking. "Contour" refers to marking regions within the image, and "Crop" involves providing the MLLM with a cropped view of the target region. The results indicate that conditioning the MLLM solely on the image or natural language fails to localize regions effectively and generate unique captions.

While previous region-level captioning models [4, 19, 31, 36, 63, 65] have demonstrated improved performance in generating unique captions when trained on URECA dataset, they lag behind URECA either because they struggle to localize multi-granularity regions, alter the original image, or overly constrain the target region without considering the global context.

This underscores that fine-tuning existing captioning models on the URECA dataset enhances their ability to handle multi-granularity captioning. However, URECA surpasses these approaches by not only generating unique captions across an image but also effectively capturing multi-granularity regions, demonstrating its capability to accurately represent regional information.

Region-level captioning. In Table 3, we present the zero-shot performance of URECA on RefCOCOg [56] and Visual Genome [22]. On RefCOCOg, URECA demonstrated competitive performance, while on Visual Genome, it achieved state-of-the-art results compared to previous approaches.

Notably, unlike prior methods, URECA achieves these results without using the benchmarks' training sets, highlighting the strong generalization ability of URECA dataset. This suggests that URECA dataset covers diverse region granularities with well-aligned captions, enabling better regional understanding. By effectively learning from a dataset with varying granularities, URECA effectively localizes and generates captions across different scales, making it highly adaptable to region-level captioning even on the zero-shot tasks.

5.2. Qualitative Results

Figure 4 shows the qualitative results comparing URECA with other methods [4, 36, 65].Unlike the comparison models, which struggle to localize regions or generate captions that capture their distinctive attributes, URECA produces unique captions for each multi-granularity region. Additional qualitative results are provided in the Appendix.

5.3. Ablation Studies

Effectiveness of mask encoder. To evaluate the effectiveness of our proposed methods, we conduct an ablation study by separately implementing each component and assessing their impact on model performance. As presented

Figure 4. **Qualitative results of the URECA and comparison models [36, 65].** Our model generates unique caption conditioned on multi-granularity regions.

Method	ROUGE	METEOR	BERTScore
Baseline	17.86	27.72	62.68
+ Mask Encoder	38.46	40.72	74.73
+ Dynamic Mask	38.95	41.25	75.11

Table 4. **Ablation study of our proposed methods on URECA dataset.**

Model Size	ROUGE	METEOR	BERTScore
1B	32.00	33.99	71.77
2B	36.64	39.00	73.92
4B	36.58	38.75	73.97
8B	38.95	41.25	75.11

Table 5. **Ablation study on model size.**

in Table 4, the baseline MLLM without conditioning performs poorly. Incorporating our mask encoder, which effectively encodes the target region while preserving its identity, significantly enhances the model's ability to localize regions and generate more descriptive captions. Furthermore, employing our dynamic masking strategy, which divides the original resolution into smaller sub-images, enables the mask encoder to capture finer details of target regions, further improving performance.

Token Length	ROUGE	METEOR	BERTScore
4	35.44	38.01	73.51
8	37.06	38.50	74.21
16	38.95	41.25	75.11

Table 6. **Ablation study on mask token length.**

MLLM size. It is well established that performance improves with larger foundation models [2, 7, 25, 62], as their knowledge capacity scales with model size. Our URECA follows this trend, achieving better performance as its size increases, as shown in Table 5. While the 1B model records the lowest performance, the largest model (8B) achieves the highest.

Mask token length. We demonstrated that our mask encoder effectively captures regions while preserving their identity. To analyze the impact of the number of tokens generated by the mask encoder, we conduct an ablation study, as shown in Table 6. We investigate the effect of increasing the number of mask tokens. As the number of tokens increases, the representation becomes more detailed, allowing for finer details to be captured, particularly in smaller regions.

6. Conclusion

We present URECA dataset, a regional captioning dataset that includes multi-granularity regions. Our primary objective is to annotate regions with unique captions that exclusively describe the target region. To achieve this, we propose an automated data pipeline that generates distinctive captions using a mask tree, which captures the hierarchical relationships between regions. To ensure high-quality evaluation, we introduce a verification stage to validate the test set. Furthermore, we introduce URECA, which encodes masked regions while effectively preserving their identity. To retain finer details, we propose dynamic masking, leveraging the LLM's flexible input length to encode masks even in high-resolution views. Our URECA achieved state-of-the-art performance on URECA dataset and demonstrates strong zero-shot captioning capabilities.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 3

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 8

[3] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 7, 12

[4] Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Vip-llava: Making large multimodal models understand arbitrary visual prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12914–12923, 2024. 2, 3, 7, 12, 14

[5] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 2, 3

[6] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2, 3

[7] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 2, 4, 7, 8, 12

[8] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024. 3

[9] Debidatta Dwibedi, Vidhi Jain, Jonathan Tompson, Andrew Zisserman, and Yusuf Aytar. Flexcap: Describe anything in images in controllable detail, 2025. 2, 3

[10] Nicola Fanelli, Gennaro Vessio, and Giovanna Castellano. I dream my painting: Connecting mllms and diffusion models via prompt generation for text-guided multi-mask inpainting. *arXiv preprint arXiv:2411.19050*, 2024. 3, 5

[11] Wentao Ge, Shunian Chen, Guiming Hardy Chen, Junying Chen, Zhihong Chen, Nuo Chen, Wenya Xie, Shuo Yan, Chenghao Zhu, Ziyue Lin, et al. Mllm-bench: evaluating multimodal llms with per-sample criteria. *arXiv preprint arXiv:2311.13951*, 2023. 5

[12] Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. Regiongpt: Towards region understanding vision language model, 2024. 2, 4, 7

[13] Shaunak Halbe, Junjiao Tian, K J Joseph, James Seale Smith, Katherine Stevo, Vineeth N Balasubramanian, and Zsolt Kira. Grounding descriptions in images informs zero-shot visual recognition, 2024. 3

[14] Jing Hao, Yuxiang Zhao, Song Chen, Yanpeng Sun, Qiang Chen, Gang Zhang, Kun Yao, Errui Ding, and Jingdong Wang. Fullanno: A data engine for enhancing image comprehension of mllms, 2024. 2

[15] Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, Qihang Yu, and Alan Yuille. Partimagenet: A large, high-quality dataset of parts. In *European Conference on Computer Vision*, pages 128–145. Springer, 2022. 2, 3, 5

[16] Miran Heo, Min-Hung Chen, De-An Huang, Sifei Liu, Subhashree Radhakrishnan, Seon Joo Kim, Yu-Chiang Frank Wang, and Ryo Hachiuma. Omni-rgpt: Unifying image and video region-level understanding via token marks, 2025. 2, 4, 7

[17] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 12

[18] Hang Hua, Qing Liu, Lingzhi Zhang, Jing Shi, Zhifei Zhang, Yilin Wang, Jianming Zhang, and Jiebo Luo. Finecaption: Compositional image captioning focusing on wherever you want at any granularity, 2024. 2

[19] Xiaoke Huang, Jianfeng Wang, Yansong Tang, Zheng Zhang, Han Hu, Jiwen Lu, Lijuan Wang, and Zicheng Liu. Segment and caption anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13405–13417, 2024. 2, 7

[20] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense caption-

ing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 12

[22] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016. 2, 3, 5, 6, 7

[23] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9579–9589, 2024. 3, 5

[24] Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11286–11315, 2024. 5

[25] Bo Li, Peiyuan Zhang, Kaichen Zhang, Fanyi Pu, Xinrun Du, Yuhao Dong, Haotian Liu, Yuanhan Zhang, Ge Zhang, Chunyuan Li, et al. Lmms-eval: Accelerating the development of large multimoal models, 2024. 8

[26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 3

[27] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. 7, 12

[28] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 2

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2, 3, 5

[30] Lixin Liu, Jiajun Tang, Xiaojun Wan, and Zongming Guo. Generating diverse and descriptive image captions using visual paraphrases. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4240–4249, 2019. 2, 4

[31] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models, 2024. 2, 3, 7

[32] OpenAI. Gpt-4o system card, 2024. 4

[33] 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. 4, 12

[34] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 7, 12

[35] Seokmok Park and Joonki Paik. Refcap: image captioning with referent objects attributes. *Scientific Reports*, 13 (1):21577, 2023. 2, 4

[36] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *ArXiv*, abs/2306, 2023. 2, 3, 7, 8

[37] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, 2016. 3

[38] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, Amir Mousavi, Yiwen Song, Abhimanyu Dubey, and Dhruv Mahajan. PACO: Parts and attributes of common objects. In *arXiv preprint arXiv:2301.01795*, 2023. 2, 3, 5

[39] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13009–13018, 2024. 2, 3, 5, 7

[40] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26374–26383, 2024. 3

[41] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms, 2023. 2, 3

[42] Yanpeng Sun, Huaxin Zhang, Qiang Chen, Xinyu Zhang, Nong Sang, Gang Zhang, Jingdong Wang, and Zechao Li. Improving multi-modal large language model through boosting vision capabilities, 2024. 2, 3, 5

[43] Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alphaclip: A clip model focusing on wherever you want, 2023. 2

[44] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 12

[45] Muntasir Wahed, Kiet A. Nguyen, Adheesh Sunil Juvekar, Xinzhuo Li, Xiaona Zhou, Vedant Shah, Tianjiao Yu, Pinar Yanardag, and Ismini Lourentzou. Prima: Multi-image vision-language models for reasoning segmentation, 2024. 2, 3, 5

[46] David Wan, Jaemin Cho, Elias Stengel-Eskin, and Mohit Bansal. Contrastive region guidance: Improving grounding in vision-language models without training. In *European Conference on Computer Vision*, pages 198–215. Springer, 2024. 7

[47] Jiuniu Wang, Wenjia Xu, Qingzhong Wang, and Antoni B Chan. Compare and reweight: Distinctive image captioning using similar images sets. In *Computer Vision–ECCV*

2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, pages 370–386. Springer, 2020. 2, 4

[48] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3

[49] Xiaoqi Wang, Wenbin He, Xiwei Xuan, Clint Sebastian, Jorge Piazentin Ono, Xin Li, Sima Behpour, Thang Doan, Liang Gou, Han-Wei Shen, et al. Use: Universal segment embeddings for open-vocabulary image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4187–4196, 2024. 2, 3, 5

[50] Zeyu Wang, Berthy Feng, Karthik Narasimhan, and Olga Russakovsky. Towards unique and informative captioning of images, 2020. 2, 4

[51] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding, 2022. 2, 3, 5, 7

[52] Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Llava-critic: Learning to evaluate multimodal models. *arXiv preprint arXiv:2410.02712*, 2024. 5

[53] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023. 5

[54] Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. Fine-grained visual prompting. *Advances in Neural Information Processing Systems*, 36:24993–25006, 2023. 4

[55] Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. Fine-grained visual prompting, 2023. 2, 3

[56] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions, 2016. 2, 3, 5, 6, 7

[57] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L. Berg. A joint speaker-listener-reinforcer model for referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[58] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. A joint speaker-listener-reinforcer model for referring expressions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7282–7290, 2017. 7

[59] Seonghoon Yu, Paul Hongsuck Seo, and Jeany Son. Pseudo-ris: Distinctive pseudo-supervision generation for referring image segmentation. In *European Conference on Computer Vision*, pages 18–36. Springer, 2024. 2

[60] Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*, 2025. 2

[61] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28202–28211, 2024. 2, 3, 5

[62] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 8

[63] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest, 2024. 2, 3, 7

[64] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 7, 12

[65] Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding, 2024. 2, 7, 8, 12, 14

[66] Yuzhong Zhao, Yue Liu, Zonghao Guo, Weijia Wu, Chen Gong, Qixiang Ye, and Fang Wan. Controlcap: Controllable region-level captioning. In *European Conference on Computer Vision*, pages 21–38. Springer, 2025. 2

[67] Li Zhou, Xu Yuan, Zenghui Sun, Zikun Zhou, and Jingsong Lan. Instruction-guided multi-granularity segmentation and captioning with large multimodal model, 2024. 3, 5

URECA 💡: Unique Region Caption Anything

Supplementary Material

A. Implementation Details

We leverage InternVL-2.5 [7] along with our mask encoder, which consists of convolutional layers followed by a two-layer MLP as the projection layer for mask tokens. For our experiments, we set the mask token length to 8. The input to the mask encoder is resized to 448×448, and the dimension of the mask tokens matches the feature dimension of the MLLM.

We train our model on four Tesla A100 GPUs (40GB) using LoRA [17]. Specifically, training is conducted in two stages: first, we train the mask encoder and projection layer, followed by LoRA fine-tuning of the MLLM. We use a batch size of 16 for LoRA tuning.

For evaluation, we adopt standard metrics used in previous studies, including BLEU [34], ROUGE [27], METEOR [3], and BERTScore [64].

B. Limitations

While our mask encoder effectively encodes multi-granularity regions without losing details, localizing the region in a sequential manner may occasionally cause the MLLM to misidentify the target region. Since we do not explicitly constrain target regions using image features or direct markers, the localization signal provided to the MLLM may be weaker compared to previous methods. Enhancing region encoding by incorporating both the mask and additional image features, rather than relying solely on sequential conditioning, could improve the MLLM's ability to accurately localize the target region.

C. More Qualitative Results

We visualize more qualitative results of URECA with previous apporaches [4, 65] in Figure A.

D. Dataset Visualization

We provide visual examples of our dataset to illustrate its diversity and complexity. Figure B showcases representative samples, highlighting key variations in object appearance, background context, and challenging scenarios. For optimal viewing, we recommend zooming in and viewing the figures in color to better observe fine details.

E. Data Pipeline

To generate unique regional captions with multi-granularity, we propose a structured four-stage process:

Stage 1: Mask Tree Construction. We first build a mask tree for each image using masks from the SA-1B dataset [21]. Intersection over Union (IoU) between masks is computed to determine containment relationships. Each tree has a root node representing the entire image, with subsequent nodes structured hierarchically based on these containment relationships.

Stage 2: Top-Down Caption Generation. In this stage, we identify primary nodes directly under the root node, termed *main objects*, whose depth exceeds a predefined threshold. Short captions are then hierarchically generated from these main objects downward through descendant nodes. Each node creates concise captions using contextual information from parent and sibling nodes to maintain coherence and uniqueness. Specific prompts used in this step are detailed in Table A.

Stage 3: Bottom-Up Caption Refinement. Short captions generated in Stage 2 are expanded into detailed descriptions. Each node enriches its caption by incorporating information from child nodes, ensuring hierarchical consistency and comprehensive context. Prompts for this refinement stage are provided in Table B.

Stage 4: Uniqueness Refinement. Finally, captions are refined by evaluating visual similarity between regions using DINO v2 [33]. Regions with high visual similarity have their captions adjusted by emphasizing distinguishing features, maintaining semantic relevance and uniqueness. Prompts for uniqueness refinement are described in Table C.

Through these stages, we systematically generate multi-granularity captions that accurately describe each region with clarity, context, and uniqueness in an automated manner.

F. Discussion

Evaluating unique caption generation for regional captioning tasks using traditional metrics such as BLEU [34], METEOR [3], ROUGE [27], and CIDEr [44] presents inherent limitations. These metrics primarily assess similarity to reference captions based on n-gram overlap, without distinguishing between essential and non-essential words. However, in unique captioning, it is crucial to generate descriptions that highlight distinctive attributes, ensuring that the caption effectively differentiates the target region from others. Existing evaluation methods treat all words

equally, failing to account for the importance of discriminative terms. As a result, captions that successfully emphasize key distinguishing features may not receive high scores if their phrasing deviates from reference texts, even if they better serve the task's objective. This limitation suggests the need for alternative evaluation approaches that better capture the quality and distinctiveness of unique captions.

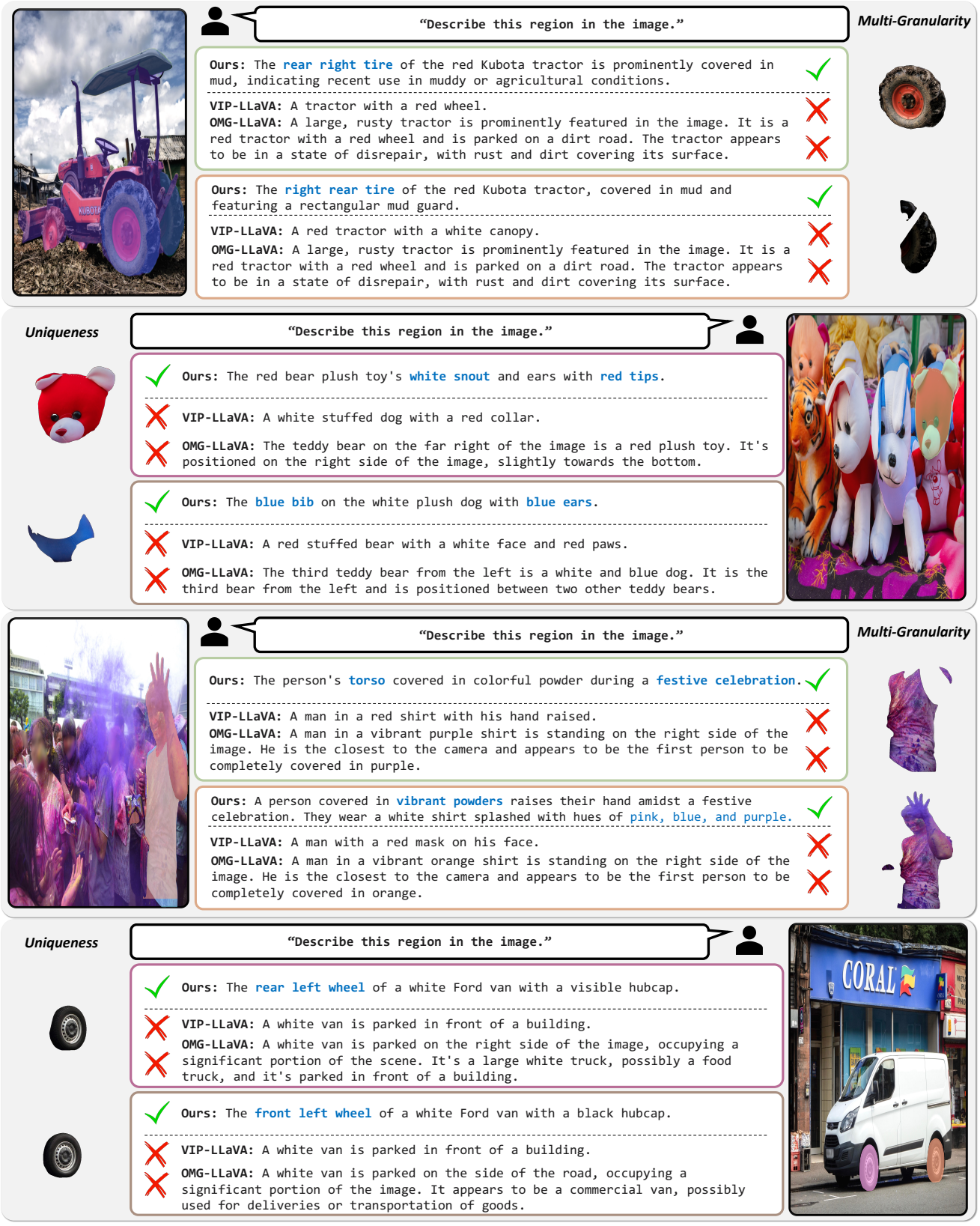

Figure A. **Qualitative results of the URECA and comparison models [4, 65].** Our model generates unique caption conditioned on multi-granularity regions.

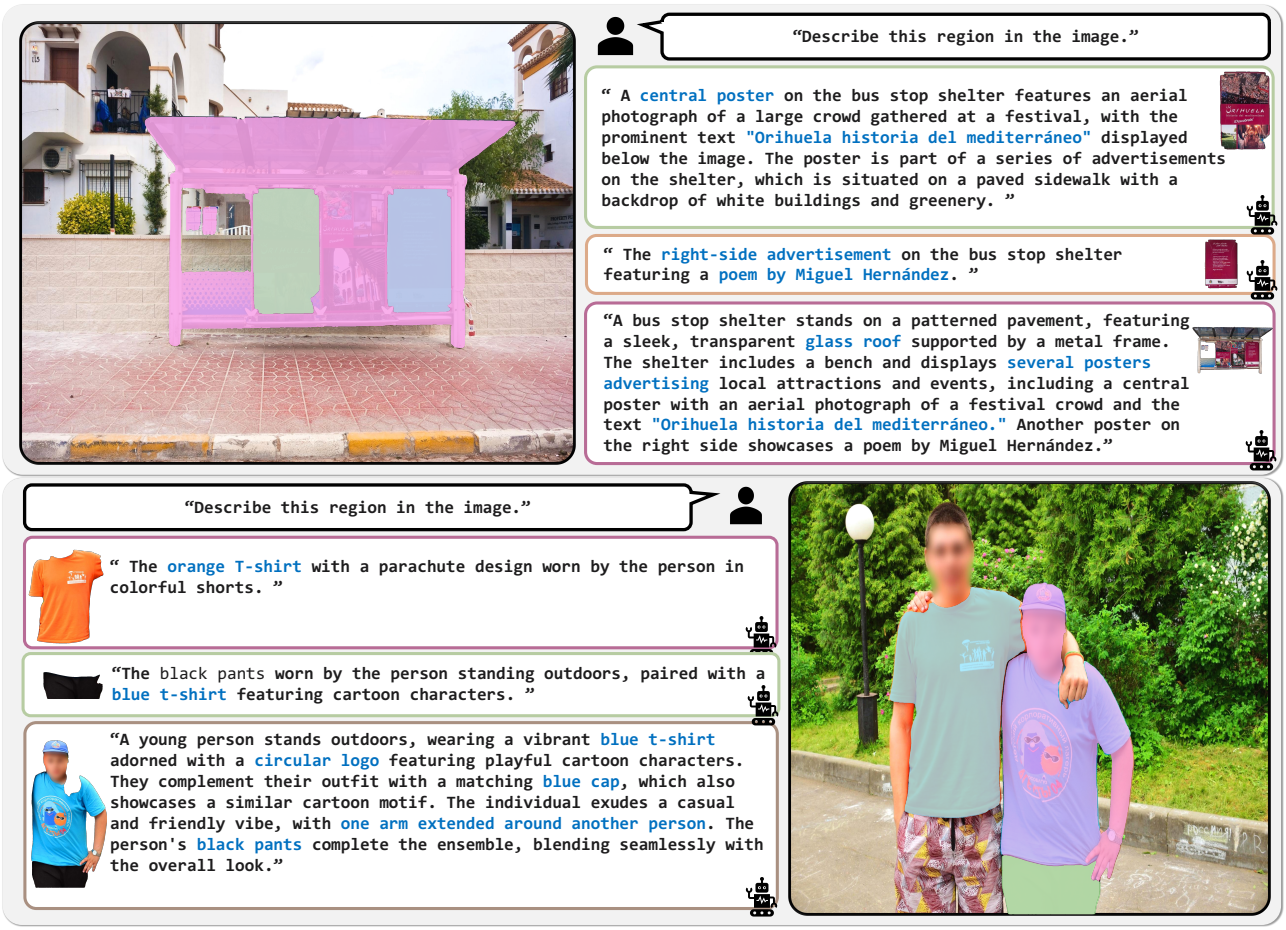

Figure B. **Example data generated by our data curation pipeline.**

```
<task>
    You are a detailed caption generator tasked with describing the main object in images.
    Your goal is to create a simple phrase that accurately represents the main object
    while avoiding hallucination.
</task>
<objectives>
    1. The main object is a subpart of a larger object; therefore, the main object alone
       may provide insufficient information.
    2. The primary focus of the caption must be on the main object while also considering
       its positional relationship or functional connection with the larger object.
    3. The primary focus of the caption must be on the main object, emphasizing attributes
       like color, texture, shape, and action if visible.
    4. The background is blurred to emphasize the main object. Focus solely on describing
       the main object in detail without mentioning the blurred background.
    5. The caption should be distinguishable from other subparts of the same larger
       object so that the region can be identified solely by looking at the caption.
       Therefore, the caption should incorporate positions or attributes that are unique
       to the main object.
    6. Creating a unique caption is important, but the most critical aspect is accuracy.
       Do not add unnecessary information solely for the sake of uniqueness.
</objectives>
<inputDetails>
    1. Image-1 highlights the main object with a yellow contour to illustrate its
       relationship with the larger object.
    2. Image-2 shows the main object cropped from the larger object.
    3. A description of the larger object will be provided in the prompt to help
       identify the main object.
    4. Descriptions of other subparts of the same larger object will also be provided.
       The caption for the main object must be clearly distinguishable from the
       descriptions of these subparts.
</inputDetails>
<descriptionOfLargerObject>
    "Description from the parent object"
</descriptionOfLargerObject>
<descriptionOfSubparts>
    "Descriptions from objects on the same level, if present."
</descriptionOfSubparts>
<outputFormat>
    1. Provide a simple phrase focusing on the main object while considering its
       positional relationship or functional connection with the larger object.
    2. The larger object may contain another object with similar attributes to the
       main object. The caption should be written in a way that clearly distinguishes the
       main object from these similar objects.
    3. Keep the caption concise, limiting it to one sentence while ensuring
       clarity and coherence.
    4. Do not explicitly mention the yellow contour or its presence in the image.
    5. Use contextual information from Image-1 to describe the main object's
       relationship with the larger object, while referencing its attributes from Image-2.
    6. Contextual details from Image-1 and the description of the larger object
       should be used only to support the description of the main object.
</outputFormat>
<outputExamples>
    "8 in-context examples"
</outputExamples>
```

Table A. Prompts for top-down generation. Captions are generated hierarchically from main objects to descendants while ensuring contextual coherence and uniqueness.

```
<task>
    You are a detailed caption generator tasked with describing the main object in images.
    Your goal is to create precise and detailed captions while avoiding hallucination.
</task>
<objectives>
    1. The caption must primarily focus on the main object while considering its
       contextual information to clearly identify what it is.
    2. The caption must emphasize the main object's attributes, such as color, texture,
       shape, and action if visible.
    3. Describe only what is visible in the image. Avoid adding any information that
       is not present.
    4. The main object is highlighted with a yellow contour.
    5. A short description of the main object will be provided in the prompt, which
       can be used to describe the main object.
    6. The main object consists of multiple subparts, and descriptions of these subparts
       will be provided in the prompt.
    7. The description of subparts may contain inaccurate, unimportant, or redundant
       information. Use only the essential details that do not contradict the
       given image to ensure that the caption for the main object compositionally
       reflects relevant information from these subparts.
</objectives>
<inputDetails>
    1. An image with the main object marked by a yellow contour will be provided.
    2. A short description of the main object will be included in the prompt.
    3. Descriptions of the subparts of the main object will also be provided in
       the prompt.
</inputDetails>
<descriptionOfMainObject>
    "Description from the main object."
</descriptionOfMainObject>
<descriptionOfSubparts>
    "Descriptions from the child objects, if present."
</descriptionOfSubparts>
<outputFormat>
    1. Provide a single descriptive paragraph that focuses on the main object.
    2. Do not use bullet points or lists.
    3. Incorporate details from the provided descriptions to accurately depict the
       main object.
    4. Never mention the presence of the yellow contour in any form.
    5. Structure the caption clearly and concisely, avoiding excessive detail or
       verbosity. Do not start with phrases like "The image shows...".
    6. Ensure the focus is evident without explicitly stating that it is the main object.
</outputFormat>
```

Table B. Prompts for bottom-up generation. Captions are refined by incorporating child node information to maintain hierarchical consistency.

```
<task>
    You are a caption refinement model that enhances given descriptions to generate unique
    and precise captions for objects in an image. Your goal is to refine the provided
    caption based on contour-based indexing while maintaining clarity and specificity.
</task>
<objectives>
    1. Describe only what is visible in the image. Avoid adding any information that is
       not present.
    2. The image contains multiple contours in different colors, each with a
       corresponding index, marking distinct objects.
    3. The main object corresponds to index 0 and is specifically outlined with a
       blue contour.
    4. Your task is to refine the caption for index 0, highlighting its unique attributes
       while clearly differentiating it from other indexed contours in the image.
    5. The refined caption must primarily focus on index 0 while considering its
       contextual information to clearly identify it from other indices.
    6. The caption must emphasize index 0's attributes, such as color, texture, shape,
       and action, to make caption unique.
</objectives>
<inputDetails>
    1. The contours in the image are color-coded, and each contour has a
       corresponding index.
    2. The index corresponding to each contour is placed at the center of the contour,
       matching its color.
    3. The initial caption for index 0 (blue contour) is provided as input.
    4. The refined caption should ensure the distinction between index 0 (blue contour)
       and other objects in the image.
</inputDetails>
<refinementGuidelines>
    1. Preserve the core meaning of the given caption while improving its specificity
       and uniqueness.
    2. Emphasize key attributes that differentiate index 0 (blue contour) from
       other indices.
    3. Avoid mentioning the presence of contours or annotations explicitly in the caption.
    4. Keep the refined caption clearly yet descriptive.
    5. Ensure that the final caption remains a natural, human-like description of
       the object.
    6. Do not use bullet points or lists.
    7. Do not start the answer with words like "Certainly!".
</refinementGuidelines>
<captionForIndex0>
    "Description from the target (index 0) object"
</captionForIndex0>
<outputFormat>
    1. Provide a single descriptive paragraph that maintains clarity and coherence
       focusing on index 0 (blue contour)
    2. The refined caption should distinguish index 0 (blue contour) from other indices.
    3. Avoid generic or ambiguous descriptions.
    4. The refined caption should make index 0 clearly stand out from the other indexed
       objects without using phrases like "distinguished by" or similar expressions.
    4. Do not reference the contour colors or indices directly.
</outputFormat>
```

Table C. Prompts for uniqueness refinement. Captions are refined by distinguishing visually similar regions while preserving semantic relevance.