URECA: Unique Region Caption Anything

Anonymous ACL submission

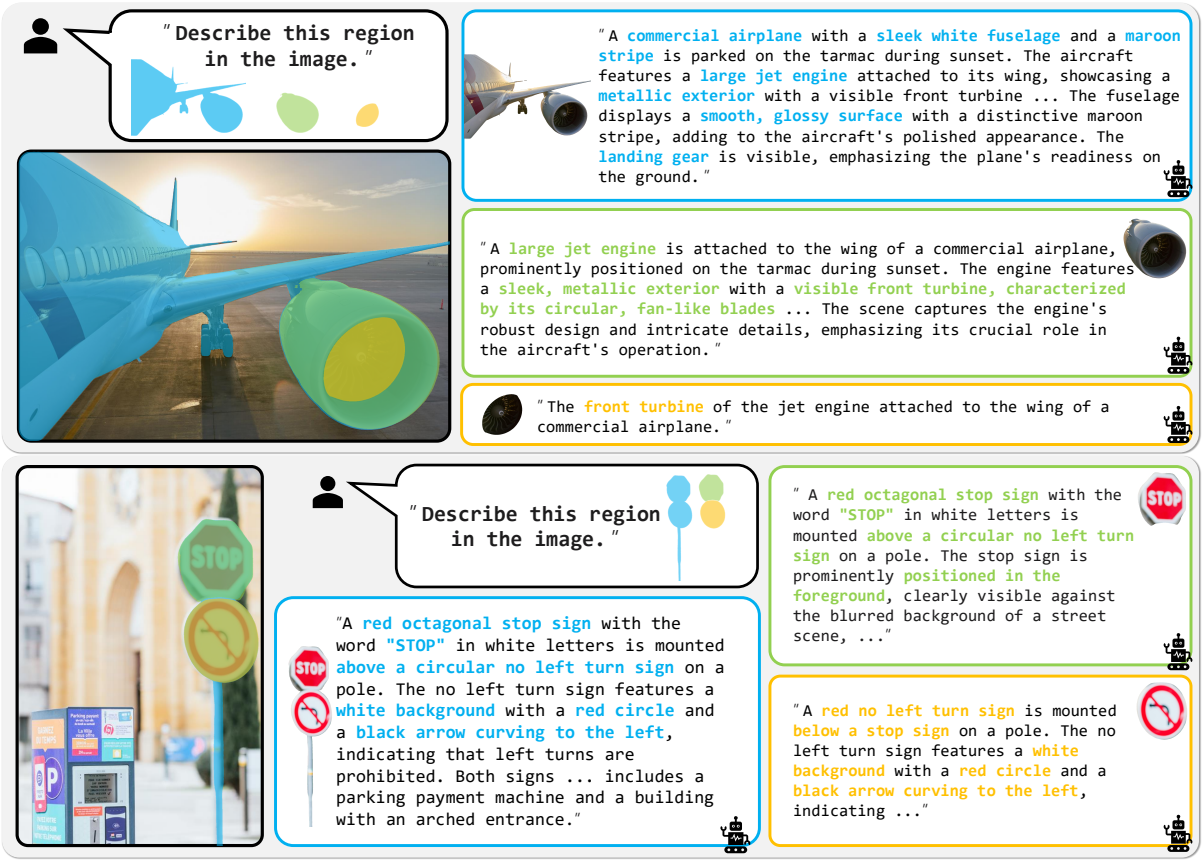

Figure 1: **Teaser.** We introduce the **Unique Region Caption Anything (URECA)** dataset, a novel region-level captioning benchmark designed to ensure caption uniqueness and support multi-granularity regions. Each caption is uniquely mapped to its corresponding region, capturing distinctive attributes that differentiate it from surrounding areas. To effectively leverage this fine-grained data, we introduce the URECA model. It features a decoupled processing strategy to preserve global context and a dynamic mask modeling technique that ensures high fidelity for regions of any scale.

Abstract

Region captioning models fail to generate descriptions that uniquely identify specific regions of interest, instead producing generic labels that could also apply to other regions within the same image. This ambiguity limits their effectiveness in downstream applications and prevents them from capturing the fine-grained details that distinguish objects. To address this, we introduce the **Unique Region Caption Anything (URECA)** dataset, a new large-scale benchmark designed to enforce caption uniqueness for multi-granularity regions. URECA dataset is constructed using a novel four-stage automated data pipeline that establishes a one-to-one mapping between a region and a descriptive caption, ensuring that each description uniquely identifies its target. We also propose the URECA model, an architecture built on two innovations for generating unique region captions: a decoupled processing strategy that preserves global context by separating region and image inputs, and dynamic mask modeling to capture fine-grained details regardless of any input image scale. Code and weights will be publicly released.

1 Introduction

Image captioning has long stood at the intersection of computer vision and natural language process-

ing, requiring models to both understand visual content and generate coherent descriptive text. Building on this foundation, region captioning presents a more challenging task: describing specific, user-defined areas within an image. While recent Vision-Language Models (VLMs) (ModelScope, 2024; Wang et al., 2025; OpenAI, 2024; Team et al., 2023) have made impressive strides in image understanding, they face a critical limitation in region captioning (Yu et al., 2016; Krishna et al., 2016; Sun et al., 2024; Yuan et al., 2024; Wu et al., 2022; Fanelli et al., 2024; Lai et al., 2024), the inability to generate unique descriptions.

Prevailing research in region captioning has largely focused on improving the fidelity of descriptions by using precise localization inputs, such as 2D coordinates, bounding boxes (Huang et al., 2024; Wang et al., 2023; Wu et al., 2022; Zhao et al., 2025), and masks (Rasheed et al., 2024), to capture fine-grained details. While these methods have achieved impressive results in generating detailed text, they often overlook a critical requirement: caption **uniqueness** within a single image. We formally define a caption as unique if it only refers to its designated region unambiguously within the context of the image, such that the description cannot be correctly applied to any other region in the same image. For example, as illustrated in Figure 1, distinct regions containing different instances of the same object class (e.g., two different women in an image) may be assigned identical, generic captions.

This failure to generate unique descriptions introduces significant ambiguity. It can cause errors in downstream applications like referring segmentation (Ding et al., 2025), which relies on a description to uniquely identify a target object. Moreover, it can confuse the model during training, as it is forced to map visually distinct inputs to identical ground-truth captions. We identify three key obstacles hindering progress:

1. **Lack of uniqueness-driven datasets.** Existing datasets (Krishna et al., 2016; Yu et al., 2016; Rasheed et al., 2024; Zhou et al., 2024) are not explicitly designed to enforce a one-to-one mapping between a region and its description. Their captions are often generic and can be reused across different instances of the same object class, thus failing to capture distinguishing visual characteristics.

2. **Poor granularity in annotations.** High-quality annotations are scarce, especially for non-salient or complex regions. Many datasets focus only on prominent objects, neglecting parts of objects, object-to-object relationships, and background elements that are crucial for comprehensive and unique descriptions.

3. **Lossy region encoding.** Despite VLM's (Chen et al., 2023; Heo et al., 2025; Lian et al., 2025) strong generative capabilities, process regional inputs in a lossy manner. Their architectures often downsample or simplify region masks, discarding the fine-grained spatial details crucial for distinguishing between similar instances. This problem is particularly severe for multi-granularity regions (e.g., small objects, thin parts), fundamentally limiting the model's ability to perceive the visual cues required for a unique caption.

To address these fundamental challenges, we introduce the **Unique Region Caption Anything (URECA)** dataset. URECA dataset is large-scale resource specifically designed to provide unique captions for multi-granularity regions. To achieve this, we developed a meticulous four-stage data pipeline that enforces a one-to-one mapping between textual descriptions and their corresponding visual areas. Unlike existing datasets that are often limited to salient objects and generic phrases, URECA dataset encompasses a diverse range of subjects including objects, parts, and backgrounds, ensuring that every caption uniquely identifies its region.

To properly evaluate a model's ability to generate captions that are both unique and accurate, we created a specialized test set with an additional verification stage to ensure data quality. Furthermore, we challenge the reliance on traditional metrics (e.g., BLEU (Papineni et al., 2002), CIDEr (Vedantam et al., 2015)), arguing that for uniqueness, semantic equivalence is more critical than the exact lexical overlap they reward. We therefore demonstrate that LLM-based evaluation metrics (Lian et al., 2025) can effectively assess semantic quality while maintaining a high correlation with traditional scores.

To leverage the fine-grained knowledge within our dataset, we propose the **URECA model**. Its architecture is founded on two key technical innovations. First, we introduce a decoupled processing strategy, where a dedicated mask encoder processes the region prompt into spatial tokens while the full image features remain unaltered. This preserves

the global context by avoiding destructive modifications to the input and precisely locate the prompt region. Second, to handle these region prompts with high fidelity across all scales, we employ a dynamic mask modeling technique that systematically tiles the mask, overcoming the fixed-input resolution limitations of visual encoders.

Experiments demonstrate that URECA significantly outperforms existing baselines, successfully interpreting region prompts to generate detailed, unique captions that are precisely grounded in the target area. By effectively resolving the ambiguity inherent in multi-granularity regions, our work establishes a new standard for discriminative region captioning. We believe that our model, dataset, and findings will serve as a foundational resource, advancing research in fine-grained visual understanding and broadly benefiting the vision-language community.

2 Related Work

Multi-modal large language models (MLLMs). MLLMs (Liu et al., 2023; Alayrac et al., 2022; Li et al., 2023) have bridged the gap between vision and language by aligning visual features with the LLM embedding space. While early works focused on holistic image understanding, research has rapidly pivoted toward fine-grained capabilities, enabling models to handle complex tasks such as reasoning over segmentation (Lai et al., 2024; Ren et al., 2024), optical character recognition (Wang et al., 2024a; Dong et al., 2024), and grounded generation (Plummer et al., 2016; Rasheed et al., 2024; Wang et al., 2024b; Zhou et al., 2024; Halbe et al., 2024). Despite these advancements, standard MLLMs often lack the specific architectural mechanisms required to process localized visual prompts with high geometric fidelity, limiting their effectiveness in dense or subtle region captioning tasks.

Region-level vision language model. Although MLLMs have demonstrated impressive image understanding capabilities, generating captions for specified regions remains a challenging task. LLaVA (Liu et al., 2023) and MiniGPT-2 (Chen et al., 2023) 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 (Cai et al., 2024; Yang et al., 2023b; Shtedritski et al., 2023) 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 (Wu et al., 2022; Dwibedi et al., 2025; Ma et al., 2024; Zhang et al., 2024a). 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 (Guo et al., 2024; Heo et al., 2025) 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 dur ing this process, resulting in the loss of meaningful features. 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

In order to generate unique caption from VLMs, high quality dataset with unique caption pair with region are crucial. To this end, we propose URECA dataset pipeline, that made with four-stage approach, enabling to build a large and diverse granularity levels with high quality unique captions.

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 (Park and Paik, 2023; Wang et al., 2020b; Liu et al., 2019; Wang et al., 2020a), 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 (Wang et al., 2020b), where models fail to generate diverse and informative captions.

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. Full implementation details for our data pipeline, including the specific prompts and parameters used to guide the annotation MLLM at each stage, are provided in the Appendix.

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 (Chen et al., 2024), for each node by referring to 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 (Yang et al., 2023c), 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 (Oquab et al., 2023). In this stage, similar-looking regions are identified using image features and marked in the image with contours and indexed bounding boxes (Yang et al., 2023a). Our annotation MLLM then generates a unique caption by explicitly differentiating the target region from other visually similar 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 (Lee et al., 2024; Xiong et al., 2024; Ge et al., 2023), 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

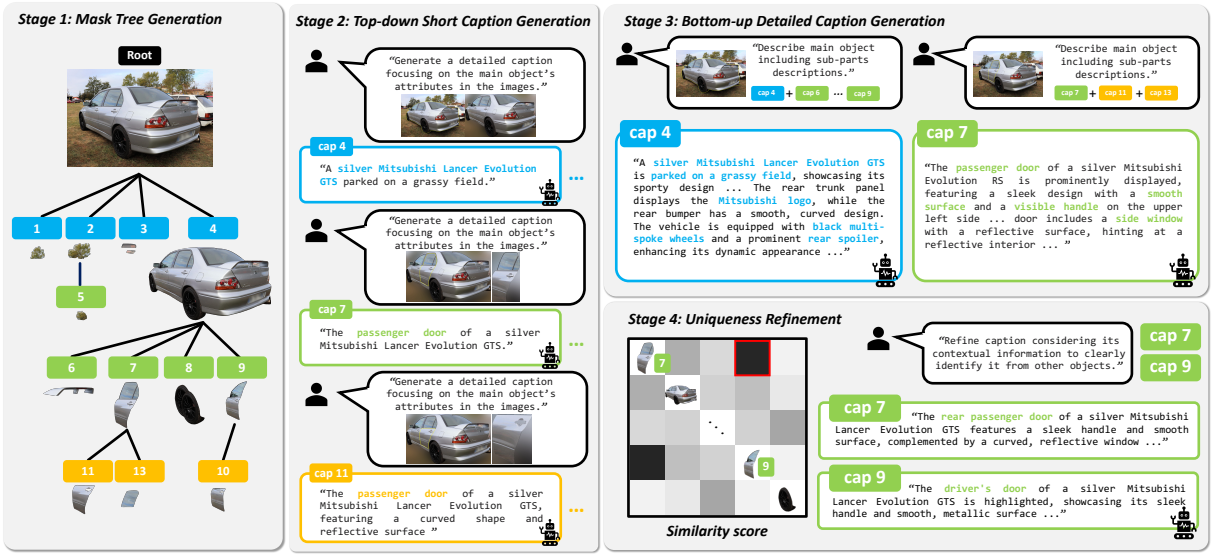

Figure 2: **Automated data curation pipeline of URECA dataset.** Our pipeline consists of four key stages to generate unique captions for multi-granularity regions. In Stage 1, we construct a mask tree that captures hierarchical relationships between regions. Stage 2 generates short captions based on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

Dataset	S	D	R	M	U	Dataset	S	D	R	M	U
RefCOCOg (Yu et al., 2016)	✓	✗	✓	✗	✗	I Dream My Painting (Fanelli et al., 2024)	✓	✓	✓	✗	✗
Visual Genome (Krishna et al., 2016)	✓	✗	✓	✗	✗	GRIT (Wu et al., 2022)	✓	✓	✓	✗	✗
PACO (Ramanathan et al., 2023)	✓	✗	✓	✗	✗	LiSA (Lai et al., 2024)	✓	✓	✓	✗	✗
PartImgNet (He et al., 2022)	✓	✗	✓	✗	✗	USE (Wang et al., 2024b)	✓	✓	✗	✓	✗
PRIMA (Wahed et al., 2024)	✓	✓	✗	✗	✗	SegCAP (Zhou et al., 2024)	✓	✓	✓	✓	✗
LLaVA-115K (Liu et al., 2023)	✓	✓	✗	✗	✗	GranD (Rasheed et al., 2024)	✓	✓	✓	✓	✗
Arcana (Sun et al., 2024)	✓	✓	✓	✗	✗	DAM (Lian et al., 2025)	✓	✓	✓	✓	✗
Osprey (Yuan et al., 2024)	✓	✓	✓	✗	✗	**URECA dataset (Ours)**	✓	✓	✓	✓	✓

Table 1: **Comparison of dataset capabilities.** **S**: Simple Caption, **D**: Dense Caption, **R**: Region Caption, **M**: Multi-granularity, **U**: Unique Caption. The proposed URECA dataset covers all captioning granularities.

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 (Ramanathan et al., 2023; He et al., 2022) that provide basic descriptions, often incorporating object classes in the captions. Dense caption represents datasets (Wahed et al., 2024; Liu et al., 2023) that include multiple attributes, offering more detailed descriptions of the region. Additionally, datasets (Yu et al., 2016; Krishna et al., 2016; Sun et al., 2024; Yuan et al., 2024; Wu et al., 2022; Fanelli et al., 2024; Lai et al., 2024) 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 (Wang et al., 2024b; Zhou et al., 2024; Rasheed et al., 2024) 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.

4 URECA Model

The overall architecture of our URECA model is illustrated in Figure 3. Its design is motivated by a central challenge in region-level understanding: how to provide a VLM with a high-fidelity representation of a specific region without compromising the global context of the full image. Existing methods for this task fall short in ways that fundamentally limit their ability to generate unique, multi-granularity captions.

The overall architecture of our URECA model is illustrated in Figure 3. Its design addresses the limitations of prior methods discussed in Section 2, specifically the loss of fine-grained spatial details

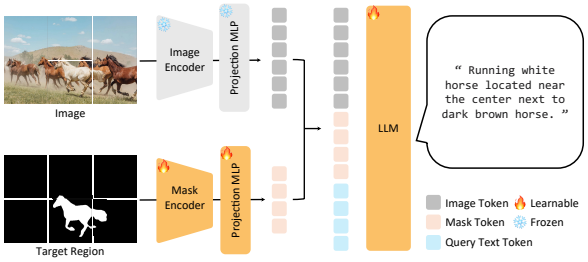

Figure 3: **Architecture of URECA model.** URECA models enables users to generate unique captions that describe distinctive attributes of any region. The mask encoder effectively encodes multi-granularity regions while preserving their identity. The mask token serves as a localizer, guiding the LLM to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

inherent in feature pooling and the visual obstruction caused by image overlays. To overcome these challenges, we introduce an architecture founded on the principle of decoupling the region's geometric information from the image's rich visual context. This approach ensures that both data streams are preserved with high fidelity.

4.1 Decoupled Processing of Region and Image Features

To generate a caption that is not just accurate but also unique, a model must understand both the specific features of a target region and its broader context within the image. Previous approaches alter the image or pooled only for corresponding region features, which irreversibly discards valuable global information and can harm the integrity of the visual features. This loss of context limits the model's ability to reason about object relationships and scene dynamics.

To overcome this limitation, we propose a decoupled processing strategy that preserves the integrity of both the image and the region prompt. Our key insight is to process the region mask and the full image in separate, parallel streams. We introduce a lightweight mask encoder that exclusively encodes the binary mask into a sequence of feature tokens. These mask tokens act as precise spatial localizers, directing the model's attention without modifying the original image.

The resulting mask tokens are then prepended to the unharmed image tokens from the vision encoder. This simple yet effective approach allows the model to leverage two complementary information sources: the mask tokens provide an unambiguous geometric cue for *where* to look, while the full image tokens provide the rich contextual information for *what* to describe. By doing so, our model effectively utilizes both local and global details to generate captions that are both spatially precise and contextually aware.

Formally, the mask encoding process is:

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

where $M \in \{0,1\}^{H \times W}$ is the input binary mask of height H and width W. The mask encoder $\phi(\cdot)$ maps $\mathbf{M}$ to a feature representation F, which consists of N spatial tokens in a D-dimensional embedding space. Unlike traditional feature pooling, our tokenization approach preserves spatial details, allowing the mask tokens to carry rich information about the region's structure. Full architectural details are provided in the Appendix.

4.2 Dynamic Mask Modeling for Multi-Granularity

To prevent the loss of fine-grained details caused by resizing high-resolution masks in fixed-input encoders, we propose dynamic mask modeling. This adaptive tiling strategy splits the original mask into a grid of sub-masks (e.g., 2×2) based on its resolution, ensuring consistent detail without excessive downsampling. Each sub-mask is encoded independently, and the resulting tokens are concatenated. Formally, the mask $M \in \{0,1\}^{H \times W}$ is divided into N_s sub-masks M_{split}:

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

This allows the mask token sequence length to scale dynamically with input resolution, preserving the subtle features required for multi-granularity captioning.

5 Experiments

5.1 Quantitative Results

We report the performance of our URECA model model on URECA test dataset. 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 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 (Chen et al.,

Model	B@1	B@2	B@3	B@4	ROUGE	METEOR	BERT-S	CLAIR
None	17.06	7.63	3.14	1.20	17.86	27.72	62.68	47.50
Contour	17.10	7.13	2.63	1.01	19.95	25.49	63.29	49.47
Crop	18.43	7.53	2.45	0.85	19.73	26.45	63.63	47.75
GPT-4o	20.38	9.01	3.62	1.53	20.44	29.87	65.44	58.62
SCA	22.76	13.58	6.97	3.88	30.76	24.87	70.67	30.82
KOSMOS-2	30.31	18.12	9.96	5.55	34.19	32.94	72.64	50.66
Osprey	31.82	20.30	12.06	7.07	36.37	34.29	73.42	53.51
OMG-LLaVA	34.01	21.88	13.51	8.46	38.14	37.29	74.68	29.09
ViP-LLaVA (7B)	34.17	22.07	13.96	9.00	38.17	37.68	74.62	55.94
ViP-LLaVA (13B)	35.35	23.52	15.07	9.96	**38.97**	39.29	74.99	55.94
URECA (Ours)	**39.29**	**23.84**	**15.42**	**9.98**	38.95	**41.25**	**75.11**	**66.96**

Table 2: **Quantitative comparison on the URECA test set.** URECA model outperforms baselines and comparison methods trained on URECA dataset across all metrics (B@n: BLEU@n, BERT-S: BERTScore).

2024) 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 (Ma et al., 2024; Huang et al., 2024; Peng et al., 2023; Zhang et al., 2024a,b; Cai et al., 2024) have demonstrated improved performance in generating unique captions when trained on URECA dataset, they lag behind URECA model 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 model 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.

Evaluation of unique captions. Traditional n-gram-based metrics are not fully equipped to evaluate caption uniqueness. A description’s uniqueness can hinge on a single discriminative word, yet conventional metrics treat all words with equal weight, failing to capture this semantic importance. To address this, recent studies have begun to adopt model-based metrics that better assess semantic meaning (Lin et al., 2025; Lian et al., 2025). We therefore provide a comprehensive evaluation using both traditional and semantic-aware metrics (Zhang et al., 2019; Chan et al., 2023), demonstrating that our model achieves state-of-the-art performance in both categories, validating its ability to generate captions that are not only accurate but also uniquely descriptive.

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 3: **Ablation study of our proposed methods on URECA dataset.**

5.2 Qualitative Results

Figure 4 provides a qualitative comparison between URECA model and baseline methods, illustrating its superior performance in handling both multi-granularity and uniqueness. In the top example, which tests multi-granularity, baseline models fail to describe the specified region, either describe it as a generic “metal bar" or hallucinating a different scene entirely. In contrast, URECA model accurately describes both the whole object (“pommel horse") and its fine-grained parts (“maroon and metallic legs"), demonstrating its precise localization and descriptive capabilities.

Similarly, in the bottom example focused on multi-granularity, where other baselines failed to locate region. URECA model, however, generates a unique caption by identifying the specific object (“the brown leather boot") and its distinguishing location (“on the man’s right foot"). This highlights our model’s ability to ground descriptions in the unique visual attributes required by the task. Additional qualitative results are provided in the Appendix.

5.3 Ablation Studies

Effectiveness of mask encoding and dynamic m asking. To evaluate the effectiveness of our proposed methods, we conduct an ablation study

Figure 4: **Qualitative results of the URECA model and comparison models (Peng et al., 2023; Zhang et al., 2024b).** Our model generates unique caption conditioned on multi-granularity regions.

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 4: **Ablation study on model size.**

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 5: **Ablation study on mask token length.**

by separately implementing each component and assessing their impact on model performance. As presented in Table 3, 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.

MLLM size. It is well established that performance improves with larger foundation models (Li et al., 2024; Chen et al., 2024; Zhang et al., 2022; Bai et al., 2025), as their knowledge capacity scales with model size. Our URECA model follows this trend, achieving better performance as its size increases, as shown in Table 4. 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 5. 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 introduce URECA dataset, a benchmark for unique, multi-granularity region captioning constructed via a novel hierarchical mask-tree pipeline. To ensure rigorous benchmarking, we include a verified test set. Complementing this, we propose URECA model, which utilizes dynamic masking to effectively encode high-resolution regions, preserving fine-grained identity and details within the LLM's flexible input space

References

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

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. 2021. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 118(41).

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

S. Balloccu and 1 others. 2025. A Survey on Data Contamination in Large Language Models. *arXiv preprint arXiv:2502.14425*.

Satanjeev Banerjee and Alon Lavie. 2005. 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.

Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. 2024. 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.

David Chan, Suzanne Petryk, Joseph E Gonzalez, Trevor Darrell, and John Canny. 2023. Clair: Evaluating image captions with large language models. *arXiv preprint arXiv:2310.12971*.

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

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

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

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, and 1 others. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Debidatta Dwibedi, Vidhi Jain, Jonathan Tompson, Andrew Zisserman, and Yusuf Aytar. 2025. Flexcap: Describe anything in images in controllable detail. *Preprint*, arXiv:2403.12026.

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

C. Fu and 1 others. 2024. MLLM-as-a-Judge: A Rigorous Benchmark for Multimodal LLM Evaluation. *arXiv preprint arXiv:2402.16741*.

Wentao Ge, Shunian Chen, Guiming Hardy Chen, Junying Chen, Zhihong Chen, Nuo Chen, Wenya Xie, Shuo Yan, Chenghao Zhu, Ziyue Lin, and 1 others. 2023. Mllm-bench: evaluating multimodal llms with per-sample criteria. *arXiv preprint arXiv:2311.13951*.

Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. 2024. Regiongpt: Towards region understanding vision language model. *Preprint*, arXiv:2403.02330.

Shaunak Halbe, Junjiao Tian, K J Joseph, James Seale Smith, Katherine Stevo, Vineeth N Balasubramanian, and Zsolt Kira. 2024. Grounding descriptions in images informs zero-shot visual recognition. *Preprint*, arXiv:2412.04429.

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

Ruozhen He, Ziyan Yang, Paola Cascante-Bonilla, Alexander C. Berg, and Vicente Ordonez. 2024. Learning from Synthetic Data for Visual Grounding. *arXiv preprint arXiv:2403.13804*.

Miran Heo, Min-Hung Chen, De-An Huang, Sifei Liu, Subhashree Radhakrishnan, Seon Joo Kim, Yu-Chiang Frank Wang, and Ryo Hachiuma. 2025. Omni-rgpt: Unifying image and video region-level understanding via token marks. *Preprint*, arXiv:2501.08326.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, and 1 others. 2022. Training Compute-Optimal Large Language Models. *arXiv preprint arXiv:2203.15556*.

Holistic AI. 2024. An Overview of Data Contamination in LLMs. https://www.holisticai.com/blog/overview-of-data-contamination.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

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

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*.

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. 2023. Segment anything. *arXiv:2304.02643*.

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. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Preprint*, arXiv:1602.07332.

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

Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. 2024. 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.

Bo Li, Peiyuan Zhang, Kaichen Zhang, Fanyi Pu, Xinrun Du, Yuhao Dong, Haotian Liu, Yuanhan Zhang, Ge Zhang, Chunyuan Li, and 1 others. 2024. Lmms-eval: Accelerating the development of large multimoal models.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Long Lian, Yifan Ding, Yunhao Ge, Sifei Liu, Hanzi Mao, Boyi Li, Marco Pavone, Ming-Yu Liu, Trevor Darrell, Adam Yala, and 1 others. 2025. Describe anything: Detailed localized image and video captioning. *arXiv preprint arXiv:2504.16072*.

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

Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. 2024. Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. *arXiv preprint arXiv:2403.20271*.

Weifeng Lin, Xinyu Wei, Ruichuan An, Tianhe Ren, Tingwei Chen, Renrui Zhang, Ziyu Guo, Wentao Zhang, Lei Zhang, and Hongsheng Li. 2025. Perceive anything: Recognize, explain, caption, and segment anything in images and videos. *arXiv preprint arXiv:2506.05302*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

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

Z. Liu and 1 others. 2024. JudgeBench: A Benchmark for Evaluating LLM-as-a-Judge. *arXiv preprint arXiv:2401.10356*.

Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. 2024. Groma: Localized visual tokenization for grounding multimodal large language models. *Preprint*, arXiv:2404.13013.

ModelScope. 2024. InternVL2.5-38B-AWQ. https://modelscope.cn/models/OpenGVLab/InternVL2_5-38B-AWQ.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, and 1 others. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

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

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

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2016. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *Preprint*, arXiv:1505.04870.

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. 2023. PACO: Parts and attributes of common objects. In *arXiv preprint arXiv:2301.01795*.

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

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

G. D. D. Sciacca and 1 others. 2025. Emergent Abilities in Large Language Models: A Survey. *arXiv preprint arXiv:2503.05788*.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. 2023. What does clip know about a red circle? visual prompt engineering for vlms. *Preprint*, arXiv:2304.06712.

Yanpeng Sun, Huaxin Zhang, Qiang Chen, Xinyu Zhang, Nong Sang, Gang Zhang, Jingdong Wang, and Zechao Li. 2024. Improving multi-modal large language model through boosting vision capabilities. *Preprint*, arXiv:2410.13733.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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

Muntasir Wahed, Kiet A. Nguyen, Adheesh Sunil Juvekar, Xinzhuo Li, Xiaona Zhou, Vedant Shah, Tianjiao Yu, Pinar Yanardag, and Ismini Lourentzou. 2024. Prima: Multi-image vision-language models for reasoning segmentation. *Preprint*, arXiv:2412.15209.

Jiuniu Wang, Wenjia Xu, Qingzhong Wang, and Antoni B Chan. 2020a. 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.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Teng Wang, Jinrui Zhang, Junjie Fei, Hao Zheng, Yunlong Tang, Zhe Li, Mingqi Gao, and Shanshan Zhao. 2023. Caption anything: Interactive image description with diverse multimodal controls. *arXiv preprint arXiv:2305.02677*.

Weiyun Wang and 1 others. 2025. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*.

Xiaoqi Wang, Wenbin He, Xiwei Xuan, Clint Sebastian, Jorge Piazentin Ono, Xin Li, Sima Behpour, Thang Doan, Liang Gou, Han-Wei Shen, and 1 others. 2024b. 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.

Zeyu Wang, Berthy Feng, Karthik Narasimhan, and Olga Russakovsky. 2020b. Towards unique and informative captioning of images. *Preprint*, arXiv:2009.03949.

Jason Wei, Yi Tay, and D. R. Tran. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

C. Wu and 1 others. 2023. Cap3D: A Captioning-Guided 3D-Text Dataset and Benchmark for Text-to-3D Generation. In *Advances in Neural Information Processing Systems*.

Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. 2022. Grit: A generative region-to-text transformer for object understanding. *Preprint*, arXiv:2212.00280.

Mingrui Wu, Xinyue Cai, Jiayi Ji, Jiale Li, Oucheng Huang, Gen Luo, Hao Fei, Guannan Jiang, Xiaoshuai Sun, and Rongrong Ji. 2024. Controlmllm: Training-free visual prompt learning for multimodal large language models. *Advances in Neural Information Processing Systems*, 37:45206–45234.

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

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

Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. 2023b. Fine-grained visual prompting. *Preprint*, arXiv:2306.04356.

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

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions. *Preprint*, arXiv:1608.00272.

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

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

Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. 2024a. Gpt4roi: Instruction tuning large language model on region-of-interest. *Preprint*, arXiv:2307.03601.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, and 1 others. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. 2024b. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *Preprint*, arXiv:2406.19389.

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

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Brooks, Eric Xing, and 1 others. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*.

Li Zhou, Xu Yuan, Zenghui Sun, Zikun Zhou, and Jingsong Lan. 2024. Instruction-guided multi-granularity segmentation and captioning with large multimodal model. *Preprint*, arXiv:2409.13407.

Jinguo Zhu and 1 others. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv preprint arXiv:2504.10479*.

Appendix

A Implementation Details

We leverage InternVL-2.5 (Chen et al., 2024) 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. While our ablation study (Table 6) indicates that performance continues to improve with 16 tokens, we selected a length of 8 to maintain a favorable balance between descriptive performance and the computational cost associated with longer token sequences during training. The input to the mask encoder is resized to 448x448, 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 (Hu et al., 2021). 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 from prior work, including BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). While these metrics allow for direct comparison, they are not designed to measure descriptive uniqueness, which is the primary goal of our research. A more detailed discussion on these limitations is provided in the Appendix G. To better assess semantic quality, we supplement these scores with BERTScore (Zhang et al., 2019) and the CLAIR score (Chan et al., 2023).

A.1 Mask Encoder Architecture

Our mask encoder is a lightweight convolutional network designed to transform a binary region mask into a sequence of feature tokens. The architecture is intentionally kept simple to ensure efficiency and reproducibility. As detailed in Algorithm 1, the encoder consists of two sequential 2D convolutional layers. Each layer uses a 3x3 kernel, a stride of 2, and padding of 1, effectively downsampling the input by a factor of 2 at each step. A ReLU activation function follows each convolution. The resulting feature map is flattened and then projected to the MLLM's hidden dimension using a two-layer MLP, which serves as the projection head. All convolutional and linear layers in the mask encoder are initialized using the Xavier normal initialization method.

Algorithm 1 Mask Encoder Pseudo-Code

Require: Binary mask $M \in \mathbb{R}^{H \times W}$
Ensure: Mask tokens $F \in \mathbb{R}^{N \times D}$ ($N = 8$, $D =$ Hidden Size)

function EncodeMask(M)
  $x \leftarrow \text{reshape}(M, [1, 1, H, W])$ ▷ Add channel dim
  ▷ First conv block: $1 \rightarrow C$
  $x \leftarrow \text{Conv2d}(x, 1, C, k = 3, s = 2, p = 1)$
  $x \leftarrow \text{ReLU}(x)$
  ▷ Second conv block: $C \rightarrow C$
  $x \leftarrow \text{Conv2d}(x, C, C, k = 3, s = 2, p = 1)$
  $x \leftarrow \text{ReLU}(x)$
  $x \leftarrow \text{flatten}(x)$ ▷ Flatten spatial dims
  ▷ Project to hidden dim
  $x \leftarrow \text{MLP}(x, \text{out} = D)$
  return x
end function

B Additional Related Work

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 (Liu et al., 2023) 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 (Alayrac et al., 2022) and BLIP-2 (Li et al., 2023) 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 (Lai et al., 2024; Ren et al., 2024), optical character recognition (Wang et al., 2024a; Dong et al., 2024), and grounding (Plummer et al., 2016; Rasheed et al., 2024; Wang et al., 2024b; Zhou et al., 2024; Halbe et al., 2024).

C Region-level Captioning

In Table 6, we present the zero-shot performance of URECA model on RefCOCOg (Yu et al., 2016) and Visual Genome (Krishna et al., 2016). On RefCOCOg, URECA model demonstrated competitive

Table 6: **Quantitative results on region-level captioning task.** Performance comparison on the METEOR for the RefCOCOg (Yu et al., 2016) and Visual Genome (Krishna et al., 2016) datasets.

Models	RefCOCOg	Visual Genome
ControlMLLM (Wu et al., 2024)	14.0	-
Kosmos-2 (Peng et al., 2023)	14.1	-
GRiT (Wu et al., 2022)	15.2	17.1
SLR (Yu et al., 2017)	15.9	-
GLaMM (Rasheed et al., 2024)	15.7	17.0
OMG-LLaVA (Zhang et al., 2024b)	15.3	-
ViP-LLaVA (Cai et al., 2024)	16.6	-
Groma (Ma et al., 2024)	16.8	16.8
RegionGPT (Guo et al., 2024)	16.9	17.0
Omni-RGPT (Heo et al., 2025)	17.0	17.0
Draw-and-Understand (Lin et al., 2024)	**23.9**	-
URECA (Zero-Shot)	16.1	**18.4**

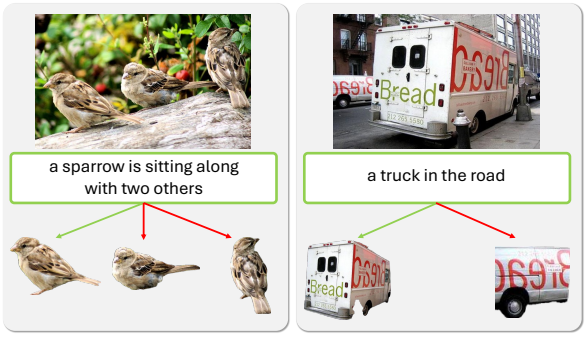

Figure 5: **Qualitative examples from the RefCOCOg dataset.** The **green arrows** indicate the *ground-truth annotation in the validation set*, while the **red arrow** highlights another *possible candidate* that can be mapped to the caption.

performance, while on Visual Genome, it achieved state-of-the-art results compared to previous approaches.

Notably, unlike prior methods, URECA model 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 model effectively localizes and generates captions across different scales, making it highly adaptable to region-level captioning even on the zero-shot tasks.

It is important, however, to acknowledge a fundamental distinction in the evaluation. As illustrated in Figure 5, datasets such as RefCOCOg and Visual Genome do not enforce unique annotations for each region. A single area—like the truck shown—can be described with a general caption ('a truck in the road') or a more specific one. This inherent ambiguity means that evaluating on these benchmarks cannot be seen as the same task as generating a single, uniquely identifying caption. Despite this misalignment, the fact that URECA model achieves such a **comparable performance** is particularly noteworthy. It underscores the model's robustness, proving its ability to generate high-quality, relevant descriptions even when the evaluation criteria are broader and less constrained than our primary objective.

D More Qualitative Results

We visualize more qualitative results of URECA model with previous apporaches (Cai et al., 2024; Zhang et al., 2024b) in Figure 6.

E Dataset Visualization

We provide visual examples of our dataset to illustrate its diversity and complexity. Figure 7 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.

F 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 (Kirillov et al., 2023). 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 7.

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 8.

Stage 4: Uniqueness Refinement. Finally, captions are refined by evaluating visual similarity between regions using DINO v2 (Oquab et al., 2023). 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 9.

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

G Discussion

Evaluating unique caption generation for regional captioning tasks using traditional metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), and CIDEr (Vedantam et al., 2015) 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.

H Limitation

Our work relies on a fully automated pipeline for dataset creation and evaluation, and as such, does not include a large-scale human study to validate the perceived quality and uniqueness of the captions. While we use GPT-4o for test set verification, which has shown strong correlation with human preferences in prior work, we acknowledge that direct human evaluation remains the gold standard. We believe this is a necessary trade-off for the scale of our dataset, and we identify rigorous human studies as a critical direction for future work.

I Methodological Justification for Dataset Curation

A potential concern regarding our dataset creation methodology is the use of two models from the InternVL family (8B for the training set, 38B for the test set), which could be perceived as an unfair evaluation setting. However, we argue that this approach is methodologically sound, does not confer an unfair advantage to our model, and aligns with state-of-the-art practices. Our justification is threefold: (1) the models are architecturally and functionally distinct, positioning the larger model as a valid "annotation oracle"; (2) the capability gap between the models is substantial and supported by established theoretical principles; and (3) the methodology aligns with broader trends in scalable, model-driven data generation.

I.1 Architectural and Training Heterogeneity

The InternVL 8B and 38B models are not merely scaled versions of one another but are heterogeneous compositions featuring significant architectural and training divergences. This compositional difference provides a strong argument against the notion of a homogenous model family.

- **Distinct LLM Backbones:** Models at different scales within the InternVL series often incorporate Large Language Model (LLM) backbones from entirely different developers. For instance, the `InternVL2.5-8B` model utilizes the `internlm2_5-7b-chat` LLM, whereas the `InternVL2.5-38B` model is built upon the `Qwen2.5-32B-Instruct` LLM (ModelScope, 2024). These LLMs are developed by separate organizations with unique architectures, training datasets, and alignment philosophies, resulting in fundamentally different internal knowledge representations and inductive biases.

- **Asymmetric Application of Advanced Training:** The larger models in the InternVL family are subjected to more advanced and qualitatively different training paradigms designed to enhance reasoning and coherence. Techniques such as Mixed Preference Optimization (MPO) and Cascade Reinforcement Learning (RL) are asymmetrically applied,

creating a significant capability gap (Wang et al., 2025; Zhu et al., 2025). For example, fine-tuning with MPO yields a 4.5-point improvement on multimodal reasoning benchmarks for the `InternVL3-38B` model, a gain attributed primarily to the training algorithm itself rather than the data (Zhu et al., 2025). This "specialized education" endows the 38B model with a more robust and human-aligned reasoning process that is qualitatively distinct from the 8B model.

I.2 Capability Gap and Oracle-Based Annotation

The architectural and training differences result in a substantial capability gap, which is consistent with established principles of AI scaling.

- **Neural Scaling Laws:** A large body of empirical research has demonstrated that model performance improves predictably as a power-law function of model parameters, dataset size, and compute (Kaplan et al., 2020; Hoffmann et al., 2022; Bahri et al., 2021). The nearly five-fold increase in parameter count from 8B to 38B is expected to yield a significant, non-linear improvement in performance, justifying the use of the larger model as a higher-quality source of ground-truth labels.

- **Emergent Abilities:** It is well-documented that capabilities can be absent in smaller-scale models but appear abruptly in larger-scale models (Wei et al., 2022; Sciacca et al., 2025). Complex, multi-step reasoning, a prerequisite for high-quality region captioning, is precisely the type of task where such emergent abilities manifest. It is therefore highly plausible that the 38B model possesses sophisticated compositional understanding and reasoning skills that are fundamentally non-existent in the 8B model.

Due to this significant capability gap, our methodology should be understood as **oracle-based annotation** rather than a form of data contamination (Balloccu et al., 2025; Holistic AI, 2024). The test set generated by the 38B "oracle" represents a target distribution of quality and complexity that a model trained on data from the much weaker 8B model cannot trivially replicate. The evaluation, therefore, remains a challenging and fair test of the model's ability to generalize towards the capabilities of a far more powerful system.

I.3 Alignment with State-of-the-Art Methodologies

Our approach follows established and peer-reviewed procedures for scalable data creation and evaluation in vision-language research.

- **The "LLM-as-a-Judge" Paradigm:** Our methodology is a logical extension of the widely accepted "LLM-as-a-Judge" framework, where powerful models like GPT-4 are used as scalable proxies for human evaluators (Zheng et al., 2023; Liu et al., 2024; Fu et al., 2024). The principle that a more capable model can reliably assess the quality of a less capable one has been validated in numerous studies, with LLM-human agreement rates often exceeding 80% (Zheng et al., 2023). If a model is trusted to *judge* quality, it can certainly be trusted to *generate* high-quality annotations.

- **Synthetic Data Generation:** The use of generative models to create training and evaluation data is a rapidly growing trend across vision-language research to overcome the bottleneck of manual annotation. Frameworks like SynGround have shown that training on purely synthetic data can significantly improve visual grounding performance (He et al., 2024). Similarly, projects like Cap3D use a pipeline of pretrained models to generate high-quality descriptive text for 3D objects at a scale that would be infeasible with human annotation (Wu et al., 2023). Our approach fits squarely within this modern paradigm of leveraging powerful foundation models for scalable dataset creation.

In summary, the significant architectural, training, and capability differences between the InternVL 8B and 38B models, combined with the alignment of our methodology with broader state-of-the-art practices, provide a robust justification for our dataset curation strategy. This approach does not create an unfair evaluation setting but rather employs a more powerful, distinct, and validated annotation oracle to establish a high-quality and challenging benchmark.

J Ethical Considerations and Potential Risks

Propagation of Bias via Automated Curation. Our dataset construction relies on an automated

pipeline utilizing large-scale foundation models, specifically InternVL2.5 (ModelScope, 2024) and GPT-4 (OpenAI, 2024), to generate and refine captions. While this enables scalability, it introduces the risk of propagating and distilling social biases present in these teacher models into our dataset. If the teacher models exhibit stereotypical associations regarding gender, race, or occupation when describing people, the URECA model may learn and amplify these biases. Although we implement a verification stage for the test set, the large-scale training set remains susceptible to inherited biases.

K Use of Large Language Models

In accordance with the ACL 2026 submission policy, we disclose that Large Language Models were used to assist in grammar correction and polishing of the writing in this paper.

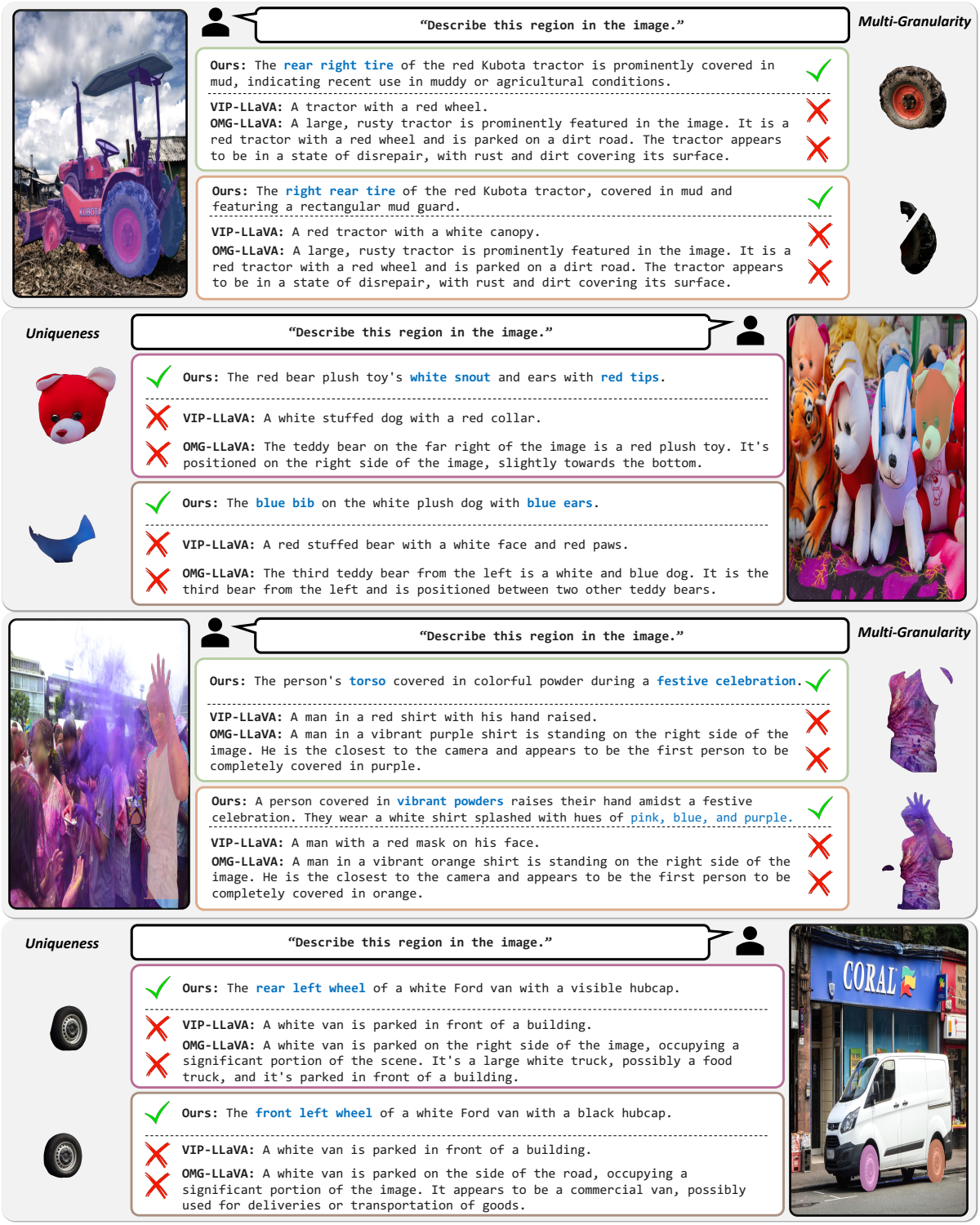

Figure 6: **Qualitative results of the URECA model and comparison models (Cai et al., 2024; Zhang et al., 2024b).** Our model generates unique caption conditioned on multi-granularity regions.

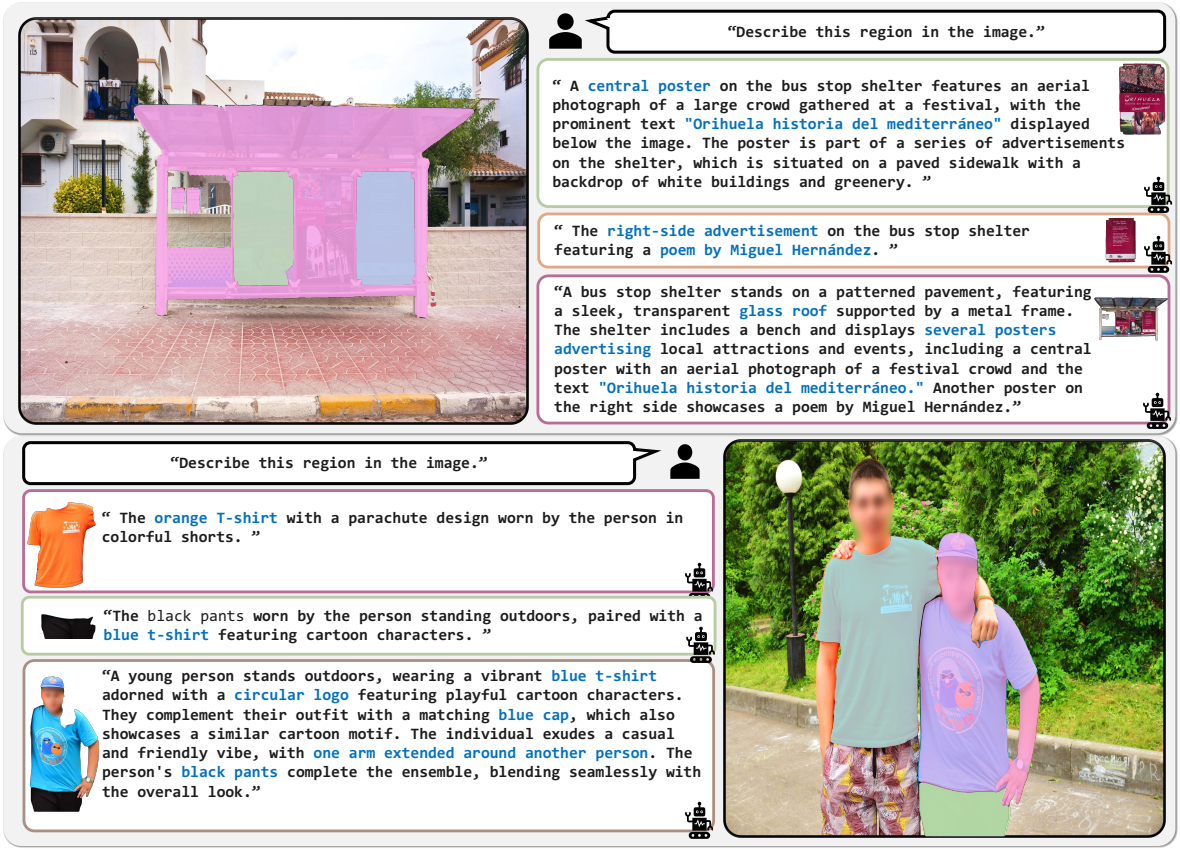

Figure 7: **Example data generated by our data curation pipeline.**

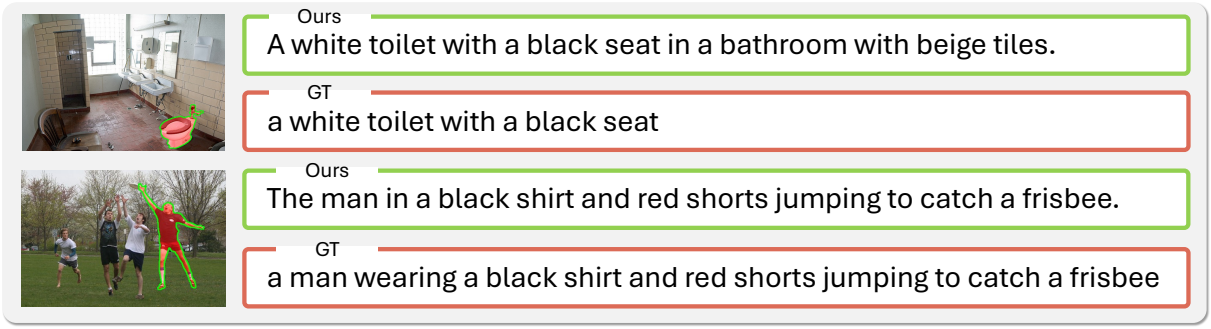

Figure 8: **Qualitative results of our URECA model on the RefCOCOg (Yu et al., 2016) dataset.**

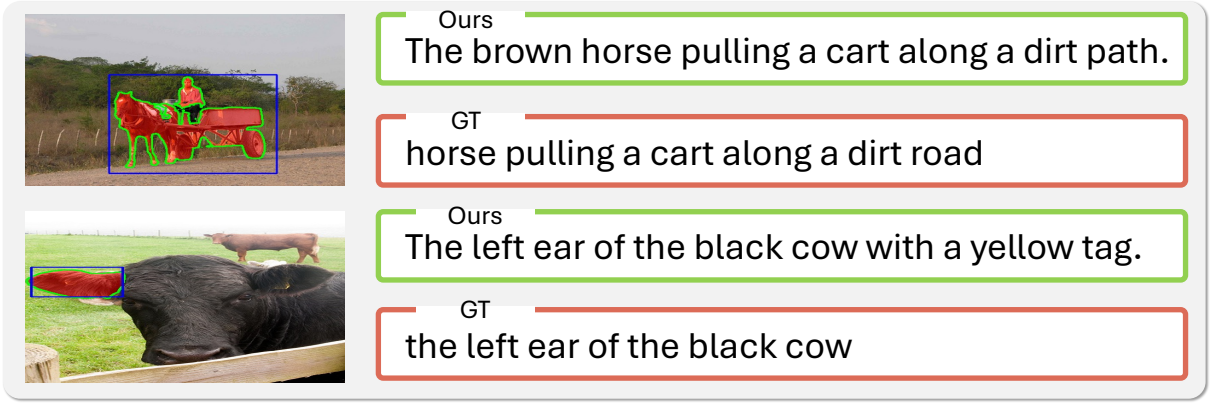

Figure 9: **Qualitative results of our URECA model on the Visual Genome (Krishna et al., 2016) dataset.**

```
<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 7: **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 8: **Prompts for bottom-up generation.** Captions are refined by incorporating child node information to maintain hierarchical consistency.

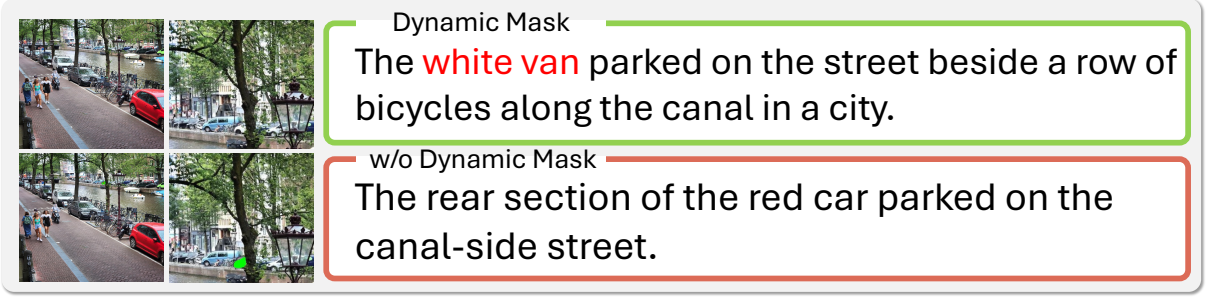

Figure 10: **Ablation study on the effect of dynamic mask.**

```
<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 9: **Prompts for uniqueness refinement.** Captions are refined by distinguishing visually similar regions while preserving semantic relevance.