

URECA: UNIQUE REGION CAPTION ANYTHING

Anonymous authors

Paper under double-blind review

ABSTRACT

Region captioning models often struggle to generate descriptions *unique* to a specific area 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.

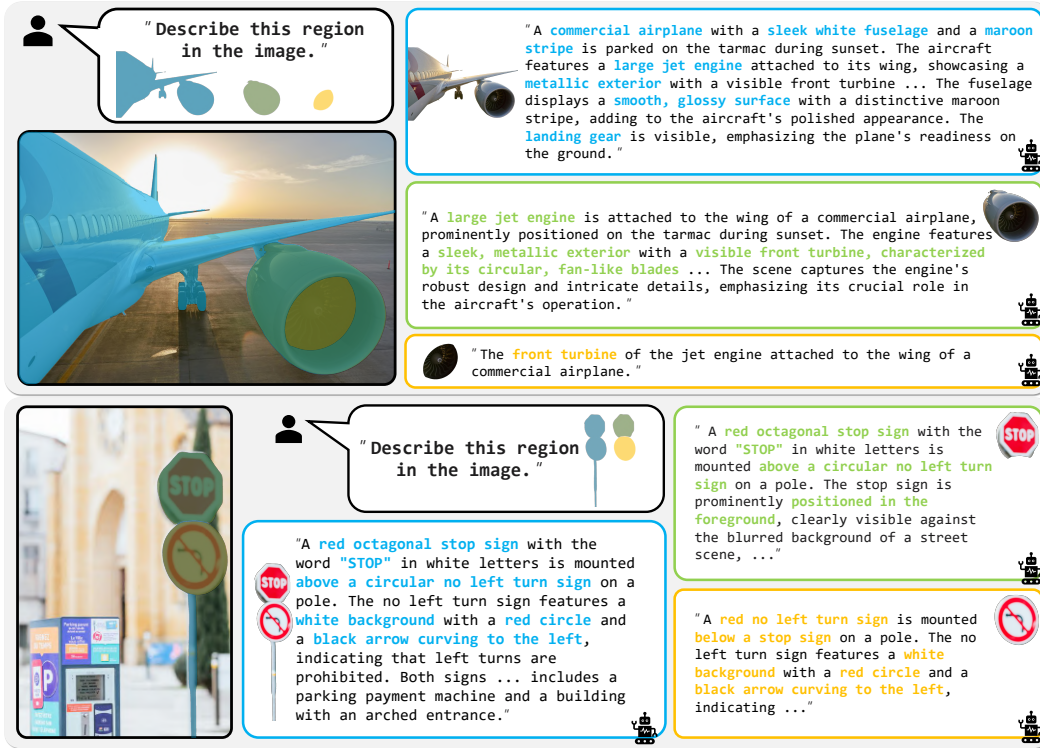


Figure 1: We introduce the **Unique Region Caption Anything (URECA)** dataset, a novel region-level captioning benchmark designed to ensure caption uniqueness and support multi-granularity regions. Each caption is uniquely mapped to its corresponding region, capturing distinctive attributes that differentiate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

1 INTRODUCTION

Image captioning is a long-standing task at the intersection of computer vision and natural language processing, requiring both a deep understanding of visual content and the ability to generate fluent descriptive text. Success hinges on the synergy between accurate visual perception and coherent language generation. Building upon this foundation, region captioning presents a more granular and complex challenge, where even state-of-the-art Vision-Language Models (VLMs) (ModelScope, 2024; Wang et al., 2025; OpenAI, 2024; Team et al., 2023) face significant open problems. Unlike image captioning, which provides a holistic but often shallow description of an entire scene, 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) demands that a model describe all relevant elements within a specific, user-defined area of interest.

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 2, 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 their strong generative capabilities, many VLMs (Chen et al., 2023; Heo et al., 2025; Lian et al., 2025) 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, a contribution largely absent from previous research. 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

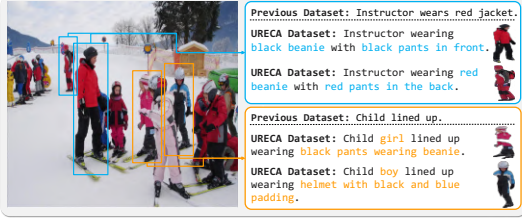


Figure 2: **Limitations of existing datasets.** This figure shows examples of region-caption pairs from widely used datasets (Krishna et al., 2016), where a single caption often applies semantically to multiple bounding boxes, resulting in a one-to-many mapping. In contrast, our URECA dataset provides unique captions for regions.

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 validate the effectiveness of our approach, demonstrating that URECA successfully interprets region prompts to generate detailed, unique captions that are precisely grounded in the target area. We believe that our model, dataset, and insights will significantly advance research in this domain and broadly benefit the vision–language community.

2 RELATED WORK

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 on regions by translating bounding box coordinates into natural language tokens. However, this approach relies heavily on the MLLM’s abstract ability to map textual coordinates to spatial locations. Other methods attempt to overlay region masks or contours directly onto the image (Cai et al., 2024; Yang et al., 2023c; Shtedritski et al., 2023). While intuitive, this permanently alters the original image, obscuring valuable contextual information.

To avoid modifying the image, another line of work performs feature pooling directly from the vision backbone’s feature maps, conditioned on either bounding boxes (Wu et al., 2022; Dwibedi et al., 2025; Ma et al., 2024; Zhang et al., 2024a) or masks (Guo et al., 2024; Heo et al., 2025). While masks provide more precise localization than ambiguous bounding boxes, the pooling operation itself introduces significant drawbacks. It is typically performed on low-resolution feature maps and aggressively aggregates spatial information, leading to a loss of fine-grained details such as shape and boundaries. In extreme cases, features for small regions can vanish entirely. Critically, all feature pooling methods share a more fundamental limitation: by extracting features only from the target region, they discard the surrounding global context. This loss of contextual information makes it nearly impossible for the model to capture the distinctiveness required for unique caption.

In summary, prior works have not adequately addressed the challenge of generating captions that are simultaneously unique, precisely localized, and applicable across multiple granularities. This gap stems from two primary factors: the lack of a large-scale dataset designed to enforce caption uniqueness, and architectures that struggle to process regional prompts without sacrificing either fine-grained detail or global context. To bridge this gap, we first propose an automated pipeline for generating a multi-granularity dataset with unique captions. We then present a novel model architecture specifically designed to leverage this data, which preserves both local attributes and global relationships, enabling it to generate truly distinctive descriptions.

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.

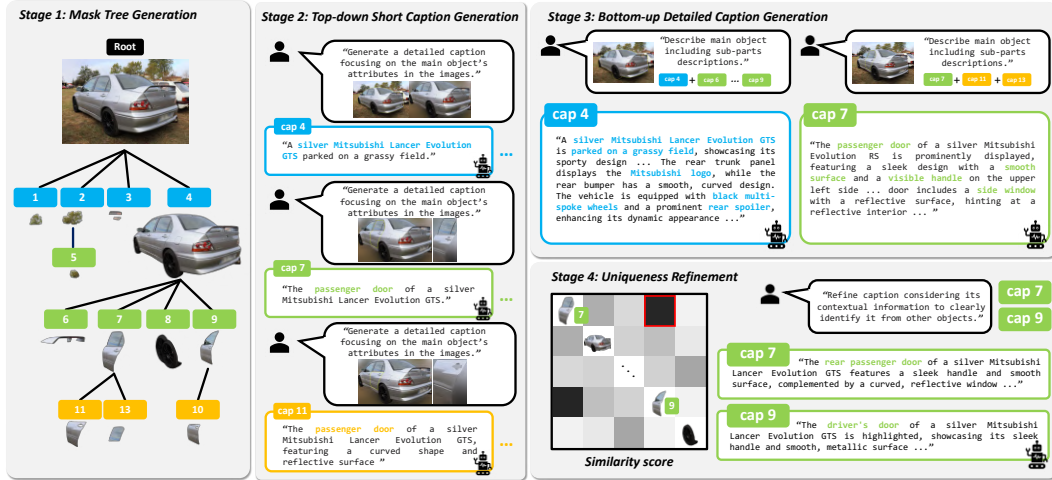


Figure 3: **Automated data curation pipeline of URECA dataset.** Our pipeline consists of four key stages to generate unique captions for multi-granularity regions. In Stage 1, we construct a mask tree that captures hierarchical relationships between regions. Stage 2 generates short captions based on 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.

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 & 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.

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 (OpenAI, 2024) 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. 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 3:

Dataset	Simple caption	Dense caption	Region caption	Multi-granularity	Unique caption
RefCOCOg (Yu et al., 2016)	✓	✗	✓	✗	✗
Visual Genome (Krishna et al., 2016)	✓	✗	✓	✗	✗
PACO (Ramanathan et al., 2023)	✓	✗	✓	✗	✗
Partimagenet (Chen et al., 2014)	✓	✗	✓	✗	✗
PRIMA (Wahed et al., 2024)	✓	✓	✗	✗	✗
LLaVA-115K (Liu et al., 2023)	✓	✓	✗	✗	✗
Arcana (Sun et al., 2024)	✓	✓	✓	✗	✗
Osprey (Yuan et al., 2024)	✓	✓	✓	✗	✗
I Dream My Painting (Fanelli et al., 2024)	✓	✓	✓	✗	✗
GRIT (Peng et al., 2023)	✓	✓	✓	✗	✗
LiSA (Lai et al., 2024)	✓	✓	✓	✗	✗
USE (Wang et al., 2024b)	✓	✓	✗	✓	✗
SegCAP (Zhou et al., 2024)	✓	✓	✓	✓	✗
Grand (Rasheed et al., 2024)	✓	✓	✓	✓	✗
DAM (Lian et al., 2025)	✓	✓	✓	✓	✗
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., (Ramanathan et al., 2023; He et al., 2022)), dense captions (e.g., (Wahed et al., 2024; Liu et al., 2023)), region captions (e.g., (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)), and multi-granularity captions (e.g., (Wang et al., 2024b; Zhou et al., 2024; Rasheed et al., 2024)). 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.

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., 2023b), 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 GPT¹, which is widely adopted to simulate human annotators for data generation tasks. Further details about the dataset pipeline can be found in Appendix.

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

¹gpt-4o-mini-2024-07-18

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 4. 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.

Previous approaches can be broadly categorized. Some encode regions using coordinate-based representations (Liu et al., 2023; Chen et al., 2023), translating bounding boxes into textual tokens. This forces the language model to learn a non-trivial mapping from text to spatial locations and struggles to represent irregularly shaped objects. Other methods rely on image modification, such as overlaying contours (Cai et al., 2024; Shtedritski et al., 2023) or providing a cropped view of the region (Cai et al., 2024). While intuitive, this alters the original visual input, potentially obscuring other objects or contextual cues that are essential for differentiating the target region from similar instances. Recent researches use feature-level pooling (e.g., RoIAlign) to extract features directly from the visual backbone’s feature maps (Wu et al., 2022; Ma et al., 2024; Rasheed et al., 2024). The critical limitation here is that pooling operates on low-resolution feature maps, leading to a significant loss of the fine-grained spatial details, such as precise boundaries and texture that are necessary to distinguish between similar objects and describe regions at varying levels of granularity.

To overcome these limitations, 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. The core components of our method, which we detail in the following subsections, are a novel processing strategy and a dynamic input mechanism designed for multi-granularity.

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

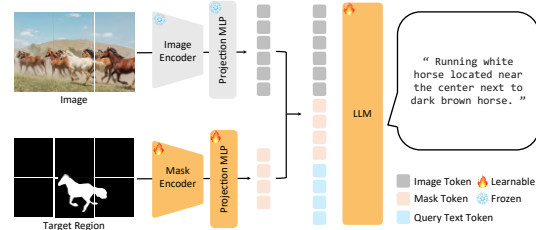


Figure 4: **Architecture of URECA model.** URECA models enables users to generate unique captions that describe distinctive attributes of any region. The mask encoder effectively encodes multi-granularity regions while preserving their identity. The mask token serves as a localizer, guiding the LLM to generate region-specific captions based on the image and query token.

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 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

A key challenge for any fixed-input encoder is handling the diverse scales inherent in multi-granularity region captioning. Resizing a large, high-resolution mask down to a small, fixed input size would inevitably lead to the loss of fine-grained details, defeating the purpose of our high-fidelity approach.

To address this, we propose a dynamic mask modeling, an adaptive tiling strategy that splits the original high-resolution mask into a grid of multiple sub-masks before encoding. Critically, the grid size (e.g., 2x2, 3x3) is adapted based on the original mask’s resolution, ensuring that each sub-mask maintains a relatively consistent and high level of detail. This prevents excessive downsampling for large regions and avoids unnecessary padding for small ones. Each sub-mask is then processed by the encoder, and the resulting sequences of tokens are concatenated.

This process splits the original mask $M \in \{0, 1\}^{H \times W}$ into multiple sub-masks M_{split} :

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

Here, N_s is the number of sub-masks in the grid. This dynamic approach allows the final mask token sequence length to scale with the input resolution, ensuring a consistently rich and detailed representation. This is particularly crucial for capturing the subtle features that distinguish small or complex objects, making it a cornerstone of URECA model’s ability to handle multi-granularity captioning.

5 EXPERIMENTS

5.1 QUANTITATIVE RESULTS

We report the performance of our URECA model on URECA dataset as well as previous benchmark datasets (Krishna et al., 2016; Yu et al., 2016). 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., 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

Table 2: **Performance comparison of URECA model with baseline methods and previous models on various evaluation metrics**, including BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee & Lavie, 2005), and BERTScore (Zhang et al., 2019). The results show that URECA model 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	BLEU@1	BLEU@2	BLEU@3	BLEU@4	ROUGE	METEOR	BERTScore	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	<u>58.62</u>
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)	<u>35.35</u>	<u>23.52</u>	<u>15.07</u>	<u>9.96</u>	38.97	<u>39.29</u>	<u>74.99</u>	55.94
URECA (Ours)	39.29	23.84	15.42	9.98	<u>38.95</u>	41.25	75.11	66.96

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.

5.2 QUALITATIVE RESULTS

Figure 5 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 masking. 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 in Table 3, the baseline MLLM without conditioning

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

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

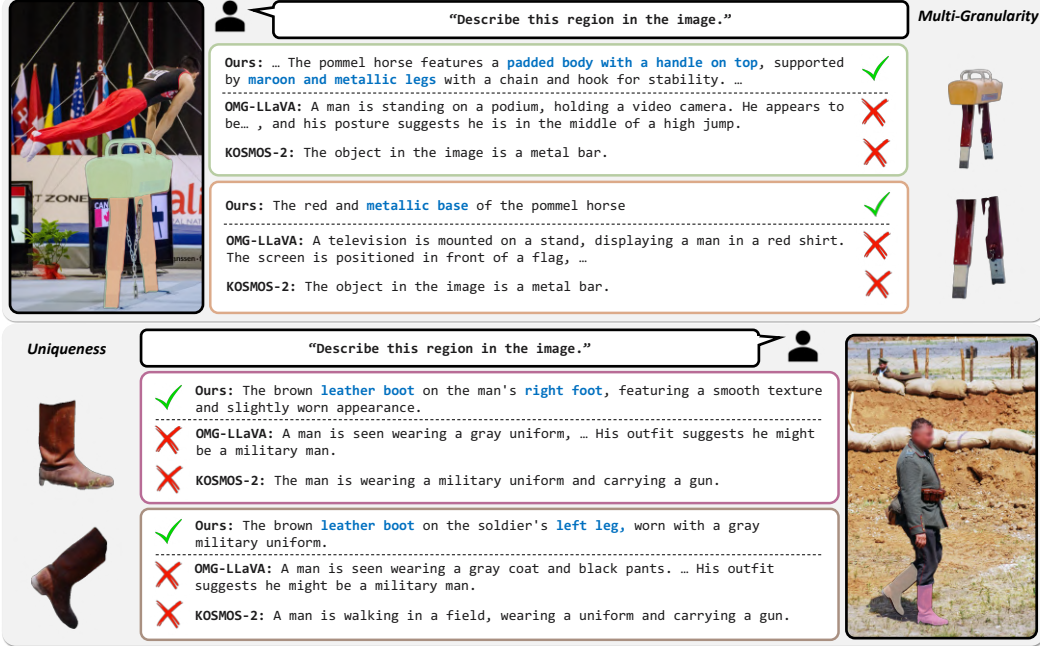


Figure 5: **Qualitative results of the URECA model and comparison models (Peng et al., 2023; Zhang et al., 2024b).** Our model generates unique caption conditioned on multi-granularity regions.

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.

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

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

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.

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

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

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 model, 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.

REPRODUCIBILITY STATEMENT

We detail the training configurations in Appendix A. We will also release our code and model checkpoints to ensure reproducibility.

REFERENCES

- 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.
- Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 118(41), 2021.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibor Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- S. Balloccu et al. A Survey on Data Contamination in Large Language Models. *arXiv preprint arXiv:2502.14425*, 2025.
- 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*, pp. 65–72, 2005.
- 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*, pp. 12914–12923, 2024.
- David Chan, Suzanne Petryk, Joseph E Gonzalez, Trevor Darrell, and John Canny. Clair: Evaluating image captions with large language models. *arXiv preprint arXiv:2310.12971*, 2023.
- Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yanyang 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.
- 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)*, June 2014.
- 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.
- Henghui Ding, Song Tang, Shuting He, Chang Liu, Zuxuan Wu, and Yu-Gang Jiang. Multimodal referring segmentation: A survey. *arXiv preprint arXiv:2508.00265*, 2025.
- 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.
- Debidatta Dwivedi, Vidhi Jain, Jonathan Tompson, Andrew Zisserman, and Yusuf Aytar. Flexcap: Describe anything in images in controllable detail, 2025. URL <https://arxiv.org/abs/2403.12026>.

- 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.
- C. Fu et al. MLLM-as-a-Judge: A Rigorous Benchmark for Multimodal LLM Evaluation. *arXiv preprint arXiv:2402.16741*, 2024.
- 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.
- 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. URL <https://arxiv.org/abs/2403.02330>.
- 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. URL <https://arxiv.org/abs/2412.04429>.
- 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*, pp. 128–145. Springer, 2022.
- Ruozhen He, Ziyang Yang, Paola Cascante-Bonilla, Alexander C. Berg, and Vicente Ordonez. Learning from Synthetic Data for Visual Grounding. *arXiv preprint arXiv:2403.13804*, 2024.
- 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. URL <https://arxiv.org/abs/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, et al. Training Compute-Optimal Large Language Models. *arXiv preprint arXiv:2203.15556*, 2022.
- Holistic AI. An Overview of Data Contamination in LLMs. <https://www.holisticai.com/blog/overview-of-data-contamination>, 2024.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL <https://arxiv.org/abs/2106.09685>.
- 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*, pp. 13405–13417, 2024.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*, 2020.
- 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.
- 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. URL <https://arxiv.org/abs/1602.07332>.
- 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)*, pp. 9579–9589, June 2024.

- 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*, pp. 11286–11315, 2024.
- 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 multimodal models, 2024.
- 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. URL <https://arxiv.org/abs/2301.12597>.
- Long Lian, Yifan Ding, Yunhao Ge, Sifei Liu, Hanzi Mao, Boyi Li, Marco Pavone, Ming-Yu Liu, Trevor Darrell, Adam Yala, et al. Describe anything: Detailed localized image and video captioning. *arXiv preprint arXiv:2504.16072*, 2025.
- Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/W04-1013>.
- Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. *arXiv preprint arXiv:2403.20271*, 2024.
- Weifeng Lin, Xinyu Wei, Ruichuan An, Tianhe Ren, Tingwei Chen, Renrui Zhang, Ziyu Guo, Wentao Zhang, Lei Zhang, and Hongsheng Li. Perceive anything: Recognize, explain, caption, and segment anything in images and videos. *arXiv preprint arXiv:2506.05302*, 2025.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.
- 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*, pp. 4240–4249, 2019.
- Z. Liu et al. JudgeBench: A Benchmark for Evaluating LLM-as-a-Judge. *arXiv preprint arXiv:2401.10356*, 2024.
- Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models, 2024. URL <https://arxiv.org/abs/2404.13013>.
- ModelScope. InternVL2.5-38B-AWQ. https://modelscope.cn/models/OpenGVLab/InternVL2_5-38B-AWQ, 2024.
- OpenAI. Gpt-4o system card, 2024. URL <https://arxiv.org/abs/2410.21276>.
- 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.
- 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*, pp. 311–318, 2002.
- Seokmok Park and Joonki Paik. Refcap: image captioning with referent objects attributes. *Scientific Reports*, 13(1):21577, 2023.
- 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.
- 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. URL <https://arxiv.org/abs/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. PACO: Parts and attributes of common objects. In *arXiv preprint arXiv:2301.01795*, 2023.
- 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)*, pp. 13009–13018, June 2024.
- 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*, pp. 26374–26383, 2024.
- G. D. D. Sciacca et al. Emergent Abilities in Large Language Models: A Survey. *arXiv preprint arXiv:2503.05788*, 2025.
- Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms, 2023. URL <https://arxiv.org/abs/2304.06712>.
- 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. URL <https://arxiv.org/abs/2410.13733>.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricutt, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- 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*, pp. 4566–4575, 2015.
- 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. URL <https://arxiv.org/abs/2412.15209>.
- 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*, pp. 370–386. Springer, 2020a.
- 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*, 2024a.
- Teng Wang, Jinrui Zhang, Junjie Fei, Hao Zheng, Yunlong Tang, Zhe Li, Mingqi Gao, and Shanshan Zhao. Caption anything: Interactive image description with diverse multimodal controls. *arXiv preprint arXiv:2305.02677*, 2023.
- Weiyun Wang et al. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*, 2025.
- Xiaoqi Wang, Wenbin He, Xiwei Xuan, Clint Sebastian, Jorge Piazzentin 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*, pp. 4187–4196, 2024b.
- Zeyu Wang, Berthy Feng, Karthik Narasimhan, and Olga Russakovsky. Towards unique and informative captioning of images, 2020b. URL <https://arxiv.org/abs/2009.03949>.
- Jason Wei, Yi Tay, and D. R. Tran. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

- C. Wu et al. Cap3D: A Captioning-Guided 3D-Text Dataset and Benchmark for Text-to-3D Generation. In *Advances in Neural Information Processing Systems*, 2023.
- 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. URL <https://arxiv.org/abs/2212.00280>.
- Mingrui Wu, Xinyue Cai, Jiayi Ji, Jiale Li, Oucheng Huang, Gen Luo, Hao Fei, Guannan Jiang, Xiaoshuai Sun, and Rongrong Ji. Controlmllm: Training-free visual prompt learning for multimodal large language models. *Advances in Neural Information Processing Systems*, 37:45206–45234, 2024.
- 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.
- 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*, 2023a.
- Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. Fine-grained visual prompting. *Advances in Neural Information Processing Systems*, 36:24993–25006, 2023b.
- Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. Fine-grained visual prompting, 2023c. URL <https://arxiv.org/abs/2306.04356>.
- Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions, 2016. URL <https://arxiv.org/abs/1608.00272>.
- 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*, pp. 7282–7290, 2017.
- 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)*, pp. 28202–28211, June 2024.
- 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, 2024a. URL <https://arxiv.org/abs/2307.03601>.
- 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.
- 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, 2024b. URL <https://arxiv.org/abs/2406.19389>.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.
- 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*, pp. 21–38. Springer, 2025.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Brooks, Eric Xing, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*, 2023.
- Li Zhou, Xu Yuan, Zenghui Sun, Zikun Zhou, and Jingsong Lan. Instruction-guided multi-granularity segmentation and captioning with large multimodal model, 2024. URL <https://arxiv.org/abs/2409.13407>.
- Jinguo Zhu et al. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv preprint arXiv:2504.10479*, 2025.

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 & 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 H. 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 of size $H \times W$

Ensure: Mask tokens F of size $N \times D$ (where $N = 8$ is the number of tokens and D is the LLM hidden size)

```

1: function ENCODEMASK( $M$ )
2:    $x \leftarrow \text{reshape}(M, [1, 1, H, W])$  ▷ Reshape mask to have a channel dimension
3:    $x \leftarrow \text{Conv2d}(x, \text{in\_channels}=1, \text{out\_channels}=C, \text{kernel}=3, \text{stride}=2, \text{padding}=1)$ 
4:    $x \leftarrow \text{ReLU}(x)$  ▷ First convolutional block
5:    $x \leftarrow \text{Conv2d}(x, \text{in\_channels}=C, \text{out\_channels}=C, \text{kernel}=3, \text{stride}=2, \text{padding}=1)$ 
6:    $x \leftarrow \text{ReLU}(x)$  ▷ Second convolutional block
7:    $x \leftarrow \text{flatten}(x)$  ▷ Flatten the spatial dimensions
8:    $x \leftarrow \text{MLP}(x, \text{out\_features}=D)$  ▷ Project to LLM hidden dim via 2-layer MLP
9:   return  $x$ 
10: 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 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.

D REGION-LEVEL CAPTIONING



Figure A: **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.

In Table A, 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 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

Table A: **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

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 A, 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.

E MORE QUALITATIVE RESULTS

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

F DATASET VISUALIZATION

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

G 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 B.

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 C.

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 D.

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

H DISCUSSION

Evaluating unique caption generation for regional captioning tasks using traditional metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee & 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.

I 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.

J 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.

J.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.

J.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.

J.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.

1026 K USE OF LARGE LANGUAGE MODELS
1027

1028 In accordance with the ICLR 2026 submission policy, we disclose that Large Language Models were
1029 used to assist in grammar correction and polishing of the writing in this paper.
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079

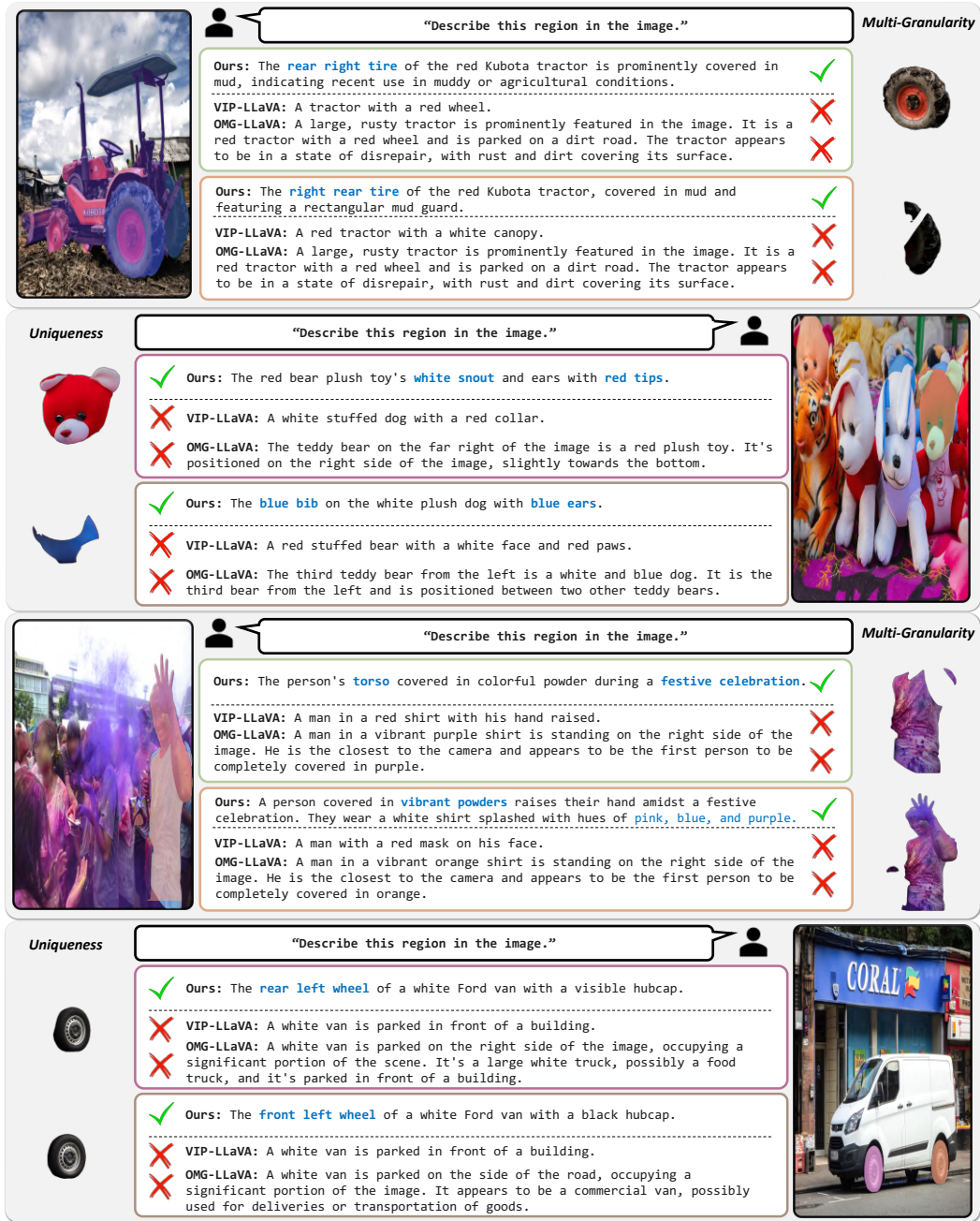


Figure B: Qualitative results of the URECA model and comparison models Cai et al. (2024); Zhang et al. (2024b). Our model generates unique caption conditioned on multi-granularity regions.

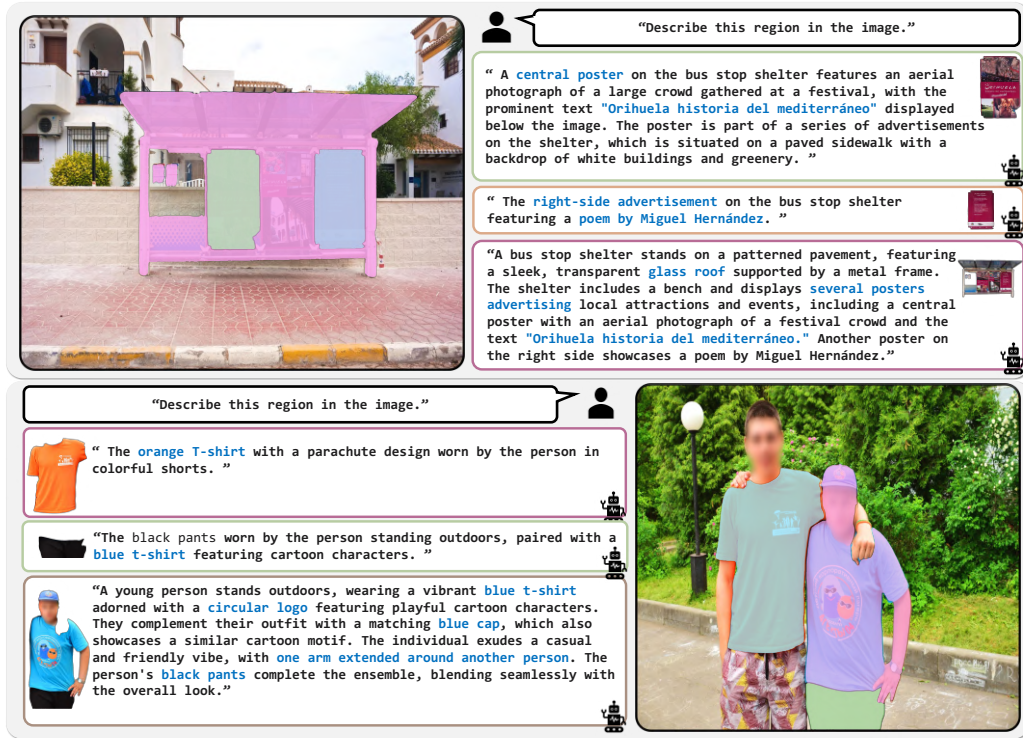


Figure C: Example data generated by our data curation pipeline.

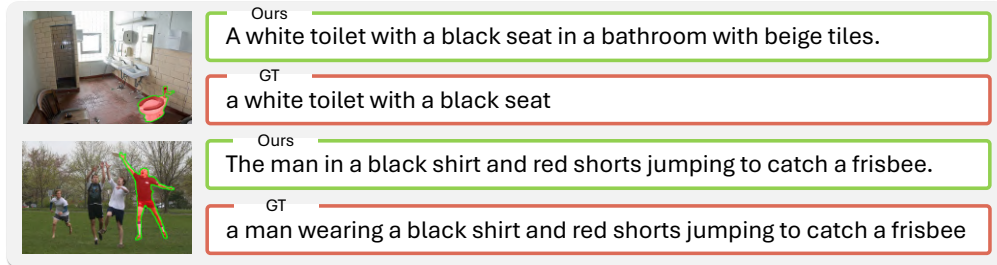


Figure D: Qualitative results of our URECA model on the RefCOCOg Yu et al. (2016) dataset.

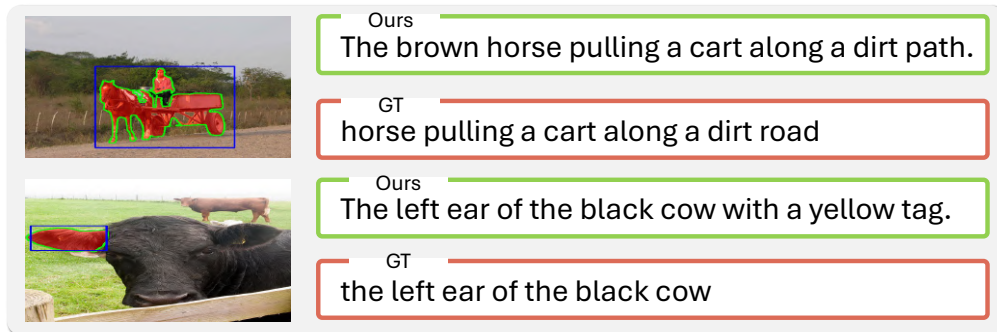


Figure E: Qualitative results of our URECA model on the Visual Genome Krishna et al. (2016) dataset.

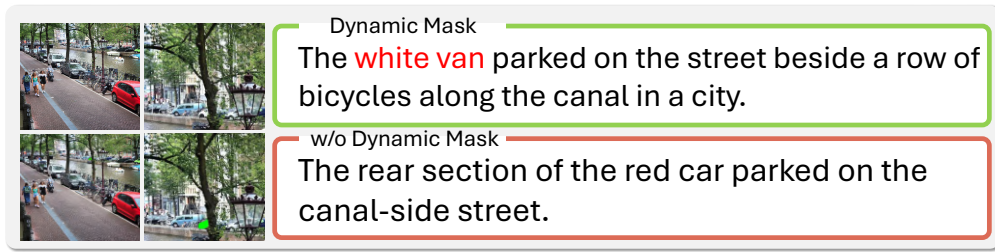


Figure F: Ablation study on the effect of dynamic mask.

```

1242 <task>
1243     You are a detailed caption generator tasked with describing the main object
1244     in images. Your goal is to create a simple phrase that accurately represents
1245     the main object while avoiding hallucination.
1246 </task>
1247 <objectives>
1248     1. The main object is a subpart of a larger object; therefore, the main
1249     object alone may provide insufficient information.
1250     2. The primary focus of the caption must be on the main object while also
1251     considering its positional relationship or functional connection with the
1252     larger object.
1253     3. The primary focus of the caption must be on the main object, emphasizing
1254     attributes like color, texture, shape, and action if visible.
1255     4. The background is blurred to emphasize the main object. Focus solely on
1256     describing the main object in detail without mentioning the blurred
1257     background.
1258     5. The caption should be distinguishable from other subparts of the same
1259     larger object so that the region can be identified solely by looking at the
1260     caption. Therefore, the caption should incorporate positions or attributes
1261     that are unique to the main object.
1262     6. Creating a unique caption is important, but the most critical aspect is
1263     accuracy. Do not add unnecessary information solely for the sake of
1264     uniqueness.
1265 </objectives>
1266 <inputDetails>
1267     1. Image-1 highlights the main object with a yellow contour to illustrate
1268     its relationship with the larger object.
1269     2. Image-2 shows the main object cropped from the larger object.
1270     3. A description of the larger object will be provided in the prompt to help
1271     identify the main object.
1272     4. Descriptions of other subparts of the same larger object will also be
1273     provided. The caption for the main object must be clearly distinguishable
1274     from the descriptions of these subparts.
1275 </inputDetails>
1276 <descriptionOfLargerObject>
1277     "Description from the parent object"
1278 </descriptionOfLargerObject>
1279 <descriptionOfSubparts>
1280     "Descriptions from objects on the same level, if present."
1281 </descriptionOfSubparts>
1282 <outputFormat>
1283     1. Provide a simple phrase focusing on the main object while considering its
1284     positional relationship or functional connection with the larger object.
1285     2. The larger object may contain another object with similar attributes to
1286     the main object. The caption should be written in a way that clearly
1287     distinguishes the main object from these similar objects.
1288     3. Keep the caption concise, limiting it to one sentence while ensuring
1289     clarity and coherence.
1290     4. Do not explicitly mention the yellow contour or its presence in the image
1291     .
1292     5. Use contextual information from Image-1 to describe the main object's
1293     relationship with the larger object, while referencing its attributes from
1294     Image-2.
1295     6. Contextual details from Image-1 and the description of the larger object
1296     should be used only to support the description of the main object.
1297 </outputFormat>
1298 <outputExamples>
1299     "8 in-context examples"
1300 </outputExamples>

```

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

```

1300 <task>
1301   You are a detailed caption generator tasked with describing the
1302   main object in images.
1303   Your goal is to create precise and detailed captions while avoiding
1304   hallucination.
1305 </task>
1306 <objectives>
1307   1. The caption must primarily focus on the main object while
1308   considering its
1309   contextual information to clearly identify what it is.
1310   2. The caption must emphasize the main object's attributes, such as
1311   color, texture, shape, and action if visible.
1312   3. Describe only what is visible in the image. Avoid adding any
1313   information that is not present.
1314   4. The main object is highlighted with a yellow contour.
1315   5. A short description of the main object will be provided in the
1316   prompt, which can be used to describe the main object.
1317   6. The main object consists of multiple subparts, and descriptions
1318   of these subparts will be provided in the prompt.
1319   7. The description of subparts may contain inaccurate, unimportant,
1320   or redundant information. Use only the essential details that do
1321   not contradict the given image to ensure that the caption for the
1322   main object compositionally reflects relevant information from these
1323   subparts.
1324 </objectives>
1325 <inputDetails>
1326   1. An image with the main object marked by a yellow contour will be
1327   provided.
1328   2. A short description of the main object will be included in the
1329   prompt.
1330   3. Descriptions of the subparts of the main object will also be
1331   provided in the prompt.
1332 </inputDetails>
1333 <descriptionOfMainObject>
1334   "Description from the main object."
1335 </descriptionOfMainObject>
1336 <descriptionOfSubparts>
1337   "Descriptions from the child objects, if present."
1338 </descriptionOfSubparts>
1339 <outputFormat>
1340   1. Provide a single descriptive paragraph that focuses on the main
1341   object.
1342   2. Do not use bullet points or lists.
1343   3. Incorporate details from the provided descriptions to accurately
1344   depict the main object.
1345   4. Never mention the presence of the yellow contour in any form.
1346   5. Structure the caption clearly and concisely, avoiding excessive
1347   detail or verbosity. Do not start with phrases like "The image shows
1348   ...".
1349   6. Ensure the focus is evident without explicitly stating that it
1350   is the main object.
1351 </outputFormat>

```

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