WeatherReasonSeg: A Benchmark for Weather-Aware Reasoning Segmentation in Visual Language Models

Wanjun Du[1*], Zifeng Yuan[2⋆], Tingting Chen[2], Fucai Ke[3], Beibei Lin[2†], and Shunli Zhang[1‡]

[1] Beijing Jiaotong University, Beijing, China
[2] National University of Singapore, Singapore
[3] Monash University, Australia
{25110603, slzhang}@bjtu.edu.cn, {zyuan, tingting.c, beibei.lin}@u.nus.edu, fucai.ke1@monash.edu

Abstract. Existing vision-language models (VLMs) have demonstrated impressive performance in reasoning-based segmentation. However, current benchmarks are primarily constructed from high-quality images captured under idealized conditions. This raises a critical question: when visual cues are severely degraded by adverse weather conditions such as rain, snow, or fog, can VLMs sustain reliable reasoning segmentation capabilities? In response to this challenge, we introduce WeatherReasonSeg, a benchmark designed to evaluate VLM performance in reasoning-based segmentation under adverse weather conditions. It consists of two complementary components. First, we construct a controllable reasoning dataset by applying synthetic weather with varying severity levels to existing segmentation datasets, enabling fine-grained robustness analysis. Second, to capture real-world complexity, we curate a real-world adverse-weather reasoning segmentation dataset with semantically consistent queries generated via mask-guided LLM prompting. We further broaden the evaluation scope across five reasoning dimensions, including functionality, application scenarios, structural attributes, interactions, and requirement matching. Extensive experiments across diverse VLMs reveal two key findings: (1) VLM performance degrades monotonically with increasing weather severity, and (2) different weather types induce distinct vulnerability patterns. We hope WeatherReasonSeg will serve as a foundation for advancing robust, weather-aware reasoning.

Keywords: Visual Language Models · Adverse Weather Robustness · Reasoning-Based Segmentation

‡ Corresponding Author
† Project Leader
⋆ Equal Contribution

Fig. 1: Adverse weather undermines VLM reasoning segmentation. Left: Under clean weather conditions, VLMs can correctly localize the target object and generate complete segmentation masks. However, when exposed to adverse weather, the model fails to accurately ground the reasoning query, resulting in incomplete segmentation masks and degraded pixel-level alignment and semantic reasoning performance. **Right:** WeatherReasonSeg addresses this limitation by introducing real-world degraded data with five structured reasoning dimensions—Function, Application, Structure, Relationship, and Requirement—to enable systematic evaluation of weather-robust reasoning.

1 Introduction

Recent advances in Vision-Language Models (VLMs) have significantly improved multimodal reasoning tasks, including reasoning-based segmentation [29, 30, 51]. Representative benchmarks such as ReasonSeg [22] enable models to generate segmentation masks conditioned on complex language descriptions rather than simple object labels. However, existing evaluations are conducted under ideal visual conditions, implicitly assuming high-quality imagery.

In real-world outdoor environments, adverse weather conditions such as fog, rain, and snow frequently degrade visual quality by reducing visibility, distorting structures, and introducing noise. These degradations disrupt visual-language alignment and can propagate through multi-stage reasoning pipelines, leading to amplified downstream errors in mask prediction. Despite its practical importance in safety-critical deployment, systematic evaluation of reasoning-based segmentation under adverse weather remains largely unexplored.

To bridge this gap, we introduce WeatherReasonSeg[†], a pioneering benchmark specifically designed to evaluate reasoning-based segmentation performance under adverse weather conditions. WeatherReasonSeg comprises two comple-

[†] `https://huggingface.co/datasets/wanwan1111/WeatherReasonSeg`

mentary components: a controllable adverse-weather dataset and a real-world adverse-weather dataset. In the former, we synthesize adverse weather effects of different types and severity levels on existing reasoning-based segmentation datasets, covering graded interference conditions including light, moderate, and severe fog, rain, and snow. This controlled setting enables systematic exploration of the reasoning limits of VLMs under varying interference intensities. Concurrently, we construct a real-world dataset to ensure the authenticity of adverse conditions, thereby enabling more accurate quantification of reasoning degradation under real visual corruption.

Furthermore, we formulate contextual queries across five critical dimensions: target function, application scenario, structural attributes, interaction relationships, and requirement matching. We adopt a mask-guided large language model prompting strategy to generate queries aligned with real images, enabling VLMs to perform context-aware reasoning and facilitating comprehensive evaluation across diverse reasoning dimensions.

Experimental results reveal a sharp deterioration in VLM performance under severe weather conditions. In synthetic benchmarks, the average accuracy of VLMs decreased by 15% compared to ideal environments; while in real-world scenarios, the accuracy of reasoning-based methods only reached half of the upper limit of perception-based methods, indicating that the reasoning stage becomes the main bottleneck. These findings quantitatively reveal the vulnerability of current VLMs to the environment and highlight the necessity of developing controllable, severity-aware benchmarks like WeatherReasonSeg to drive the development of weather-robust reasoning models.

Our contributions are summarized as follows:

- We propose the first benchmark, **WeatherReasonSeg**, specifically designed to evaluate the reasoning capabilities of VLMs under adverse weather conditions, bridging the gap between idealized benchmark evaluation and real-world deployment scenarios.
- We construct a synthetic dataset with graded weather interference, covering multiple weather types and severity levels, which quantitatively characterizes the impact of weather intensity on reasoning processes and segmentation performance.
- We build a real-world reasoning-based segmentation dataset and adopt a mask-guided large language model prompting strategy to generate semantically consistent, executable, and highly diverse queries. These queries encourage VLMs to perform complex compositional reasoning, extending beyond the task scope of traditional reasoning-based segmentation.

2 Related Work

Reasoning in Large Models Recent years have witnessed substantial advances in the reasoning capabilities of Large Language Models (LLMs) [17]. OpenAI-o1 [35] shows that extending Chain-of-Thought (CoT) reasoning during inference, which is often referred to as inference-time scaling, can significantly

improve complex reasoning. Building upon this paradigm, subsequent works explore test-time scaling strategies via process-based reward modeling [26, 50, 53], reinforcement learning (RL) optimization [20, 28, 45], and search-based inference mechanisms [9, 49]. Among these approaches, DeepSeek-R1 [10] employs the GRPO algorithm [45] and achieves strong reasoning performance with only a few thousand RL training steps, highlighting the efficiency of RL-based policy optimization for enhancing reasoning ability.

Inspired by these advances in LLM reasoning, recent efforts have begun transferring reasoning-oriented training paradigms to multimodal large language models (MLLMs). For example, Open-R1-Multimodal [21] focuses on multimodal mathematical reasoning, while R1-V [47] demonstrates improvements in visual counting tasks. However, existing multimodal reasoning methods (*e.g.*, [15, 16, 18, 21, 46, 48]) primarily focus on high-level semantic or symbolic reasoning and are typically evaluated on benchmarks such as OK-VQA [34], A-OKVQA [44], and VCR [62], which involve clear and well-curated images. As a result, these approaches seldom address fine-grained pixel-level understanding or reasoning under visually degraded conditions. In contrast, our work concentrates on pixel-level reasoning for visual perception and investigates reinforcement learning as a principled mechanism to bridge reasoning and dense segmentation.

Semantic Segmentation with Reasoning Semantic segmentation aims to perform dense pixel-wise classification by assigning a semantic category label to each pixel. Extensive prior research [1, 3, 4, 8, 27, 31, 41, 63], has driven significant progress in this area. Representative models such as DeepLab [5], MaskFormer [7], and Segment Anything Model (SAM) [19, 38] establish strong and stable baselines, making conventional category-driven segmentation a relatively mature problem.

To overcome the limitations of predefined label spaces, referring expression segmentation (RES) [14, 61] introduces natural language descriptions to specify target regions. In this setting, models must align short textual expressions with corresponding visual entities. LISA [22] further extends this framework to reasoning-based segmentation, where queries may be longer, more abstract, or require multi-step reasoning. Such tasks demand deeper joint reasoning over linguistic semantics and visual cues to accurately localize and segment the intended object.

MLLMs for Segmentation Following the introduction of the <SEG> token in LISA [22, 59], which enables interaction between multimodal large language models (MLLMs) and segmentation architectures, a number of subsequent works [2, 6, 40] explore integrating MLLMs into segmentation pipelines. Many of these methods, including OneTokenSegAll [2] and PixelLM [40], adopt token-based designs that connect language models with segmentation decoders via special interface tokens.

While effective, such tightly coupled frameworks typically require large-scale annotated data to jointly fine-tune both the MLLM and the segmentation module. Moreover, additional token-level interactions may disturb the original pixel-

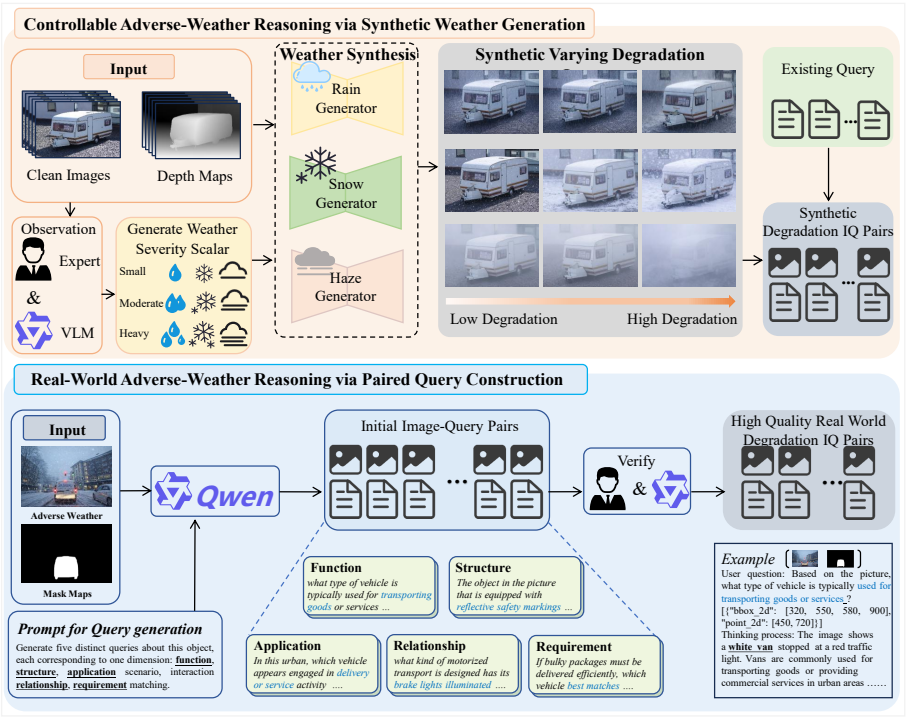

Fig. 2: Overview of WeatherReasonSeg. The framework consists of two complementary components. **Top:** A controllable synthetic weather generation process can generate weather type degradations (rain, snow, haze) of varying severity to construct synthetic image-query pairs for robustness evaluation. **Bottom:** A real-world adverse-weather reasoning dataset constructed via mask-guided large language model prompting, followed by human–model collaborative verification to ensure semantic alignment and reasoning validity.

level representations learned by pretrained segmentation networks. In contrast, our approach adopts a decoupled design that preserves the structural integrity of existing segmentation models while leveraging the reasoning capability of MLLMs to enhance segmentation performance.

3 Methodology

Figure 2 illustrates the overall pipeline of WeatherReasonSeg, which comprises two complementary components: a controllable adverse-weather reasoning dataset and a real-world adverse-weather reasoning dataset. The first dataset synthesizes different weather degradations across three severity levels based on existing reasoning datasets [22], providing a controlled environment to evaluate model stability under varying conditions. Complementarily, the second dataset

leverages a mask-guided LLM prompting to construct high-quality, real-world image-query pairs, capturing the diverse reasoning dimensions inherent in actual adverse weather.

3.1 Controllable Adverse-Weather Reasoning via Synthetic Weather Generation

To enable a systematic robustness evaluation of VLM reasoning capabilities, we adopt an existing reasoning-based segmentation dataset as the foundation and synthesize controllable weather degradations on top of it. This synthetic strategy allows explicit manipulation of weather types and severity levels (e.g., fog density, rainfall intensity, and snow accumulation), thereby supporting fine-grained analysis of how progressively increasing visual disturbances affect reasoning stability.

Base Dataset ReasonSeg [22] is the first and most representative benchmark specifically designed for reasoning-driven segmentation. Unlike conventional semantic segmentation or referring expression datasets, ReasonSeg emphasizes implicit reasoning without relying on explicit category labels, requiring models to infer targets through attributes, relationships, and contextual constraints.

Adverse Weather Synthesis Process Building upon the need for controllable degradation modeling, we introduce a physically interpretable weather synthesis mechanism into the ReasonSeg dataset, forming the synthetic component of our benchmark-WeatherReasonSeg. The design follows two core principles. First, the synthesis process should follow real imaging physics rather than simple noise injection, which ensures realistic degradation effects. Second, the degradation severity must be continuously controllable. This allows systematic analysis of model performance under different weather types and levels. The overall pipeline is illustrated in Figure 2. Clean images and their corresponding depth maps are fed into dedicated weather generators for rain, snow, and fog. Based on a weather severity scalar derived from expert knowledge and VLM-assisted calibration, different levels of degradation are applied to the images. The degraded images are then paired with the original reasoning queries to construct degraded image–query pairs.

Specifically, we model three representative weather types, namely rain, snow, and fog, according to their underlying physical characteristics, incorporating effects such as rain streak accumulation, snowflake scattering and occlusion, and atmospheric light scattering. The severity of each weather type is continuously controlled by adjusting physically meaningful parameters, such as rain density, snow particle distribution, and visibility distance, thereby generating degradations ranging from light to severe conditions. Overall, this physics-guided and controllable synthesis framework enables structured and fine-grained modeling of adverse weather across multiple types and severity levels, facilitating systematic analysis of how environmental disturbances influence VLM reasoning stability. Detailed implementation of the synthesis process and specific parameter settings for each weather generator are provided in the Supplementary Material.

Table 1: Comparison of recent benchmarks and resources for vision-language reasoning and segmentation. We evaluate each dataset by whether it provides bounding boxes and segmentation masks, supports general reasoning and adverse-weather settings, and by its data type and amount.

Dataset (Venue, Year)	Bounding Boxes	Seg. Masks	General Reasoning	Adverse Weather	Data Type	Amount
RainCityscapes (CVPR'18) [12]	✗	✓	✗	✓	Syn	10.6k
ACDC (ICCV'21) [43]	✓	✓	✗	✓	Real	4k
Rain Wcity (IJCAI'22) [64]	✗	✓	✗	✓	Real	24.3k
CREPE (CVPR'23) [33]	✓	✗	✓	✗	–	1162.6k
LVIS-Ground (ECCV'24) [32]	✗	✗	✓	✗	–	4.3k
M4-Instruct (CVPR'24) [24]	✗	✗	✓	✗	–	307k
FP-RefCOCO (CVPR'24) [57]	✓	✓	✗	✗	–	65.8k
LLM-Seg40K (CVPR'24) [52]	✓	✓	✗	✗	–	40k
MRES-32M (CVPR'24) [55]	✓	✓	✗	✗	–	32M
NaturalBench (CVPR'24) [23]	✗	✗	✓	✗	–	10k
VAB (CVPR'25) [11]	✗	✗	✓	✗	–	0.54k
ReasonSeg (CVPR'24) [22]	✓	✓	✓	✗	–	1.2k
R-RIS (TIP'24) [56]	✓	✓	✗	✗	–	42k
HalluSegBench (arXiv'25) [25]	✗	✓	✓	✗	–	3.7k
WeatherReasonSeg (ours)	✓	✓	✓	✓	Syn & Real	**44.7k**

3.2 Real-World Adverse-Weather Reasoning via Paired Query Construction

Although physics-based synthetic data can systematically control the type and severity of image degradation, it cannot fully capture the complexity of real-world conditions. Therefore, we further construct a reasoning-based segmentation dataset for real-world adverse weather conditions. This dataset uses real-world degraded images while preserving high-quality pixel-level annotations. Based on this, we employ a mask-guided large-scale language model prompting mechanism to generate multidimensional natural language reasoning queries.

Base Dataset Firstly, we adopt ACDC (Adverse Conditions Dataset with Correspondences) [42] as the foundation for real-world degraded scenarios. It covers multiple adverse weather conditions, including rain, fog, snow, and low-light (nighttime) environments. A key advantage of ACDC is its high-quality pixel-level annotations and cross-weather consistency, making it well-suited for studying the impact of visual degradation on perception and reasoning. Unlike synthetic data, the degradations arise from real physical environments and imaging processes, including complex occlusions, non-uniform scattering, and changes in semantic visibility.

Based on the original segmentation annotations, we extract target masks and construct reasoning-driven language queries aligned with the pixel-level regions, extending ACDC into a real-world degraded dataset tailored for visual-language reasoning segmentation tasks.

Query Generation To systematically evaluate the reasoning capabilities of VLMs in realistic degradation scenarios, we employ a mask-guided prompting strategy to automatically generate natural language queries that are strictly

aligned with the target segmentation region. Furthermore, we pose questions about the same target object from multiple reasoning perspectives to characterize the differences in VLM's reasoning behavior across different contextual dimensions. As illustrated in Figure 2, adverse weather images and the corresponding extracted mask maps are input into a large language model. Guided by carefully designed prompts, the model generates queries across five different reasoning dimensions, forming initial image-query pairs. These pairs are then jointly validated and filtered by human annotators and the large language model according to predefined rules. The final output consists of high-quality real-world degradation image-query pairs.

Unlike queries that only describe object categories or simple attributes [22], we design queries from five key reasoning dimensions [13, 36, 40, 51, 54, 58]: (1) *Function* focuses on the intrinsic purpose or core affordance of the target object, independent of a specific user demand. For example, a traffic light is used to regulate traffic flow, and a truck is used to transport goods. (2) *Application Scenario* emphasizes the contextual usage environment of the target object, such as a bus operating in public transportation or a service vehicle working in an urban maintenance scenario. (3) *Structural/Feature* targets directly observable visual attributes and compositional details, such as the cargo compartment of a truck, the sliding doors of a bus, or the three-color signal structure of a traffic light. (4) *Relational* examines the spatial or interactive relationships between the target object and surrounding entities, such as the relationship between pedestrians and roads, buses and bus stops, or vehicles and traffic infrastructure. (5) *Requirement Matching* maps practical needs or task constraints to appropriate target objects. For example, when bulky goods need to be transported, the expected answer may be a truck; when one needs to take public transportation, the target may be a bus or bus stop.

We note that *Function* and *Requirement Matching* are related but distinct: the former is object-centric and focuses on an object's intrinsic purpose, while the latter is goal-centric and maps practical needs to suitable visual entities.Illustrative examples of these five reasoning dimensions are provided in Figures 1 and 2.

Query Dataset Filtering Beyond careful image selection and prompt design, we implement a rigorous filtering strategy to ensure query quality and semantic consistency. This process eliminates query–answer pairs with potential ambiguity or task misalignment. Specifically: (a) We discard questions that merely enumerate objects or components without functional or contextual semantics, avoiding shallow descriptive queries. (b) We remove question groups that fail to maintain single-target consistency, where all five questions do not explicitly refer to the same physical object. (c) We retain only question–answer groups in which all questions refer to the same explicit and visually identifiable named entity, ensuring semantic clarity and disambiguation. (d) We exclude question–answer pairs that mention objects absent from the image, visually unclear, or semantically ambiguous, ensuring all queries can be reliably answered based solely on the input image.

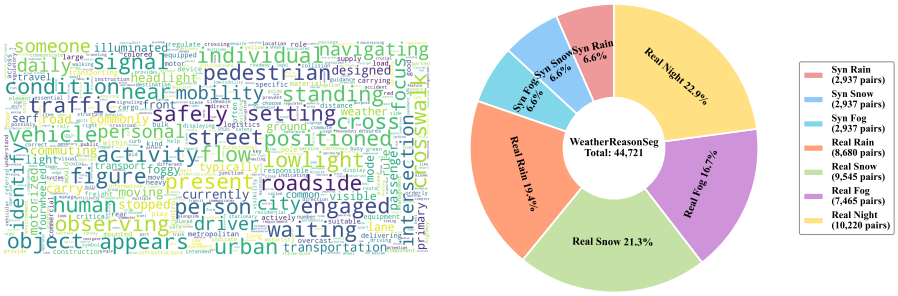

Fig. 3: Overview of WeatherReasonSeg. Left: Query word cloud. Right: Distribution of query–image pairs across weather conditions.

3.3 Dataset Statistics and Analysis

Our WeatherReasonSeg contains a total of 44,721 image-query pairs spanning both synthetic and real-world environments. The synthetic subset includes 2,937 image–query pairs for each weather type (rain, snow, and fog), amounting to 8,811 pairs for controlled robustness evaluation. The real-world subset further enhances ecological validity, comprising 8,680 rainy, 9,545 snowy, 7,465 foggy, and 10,220 nighttime pairs, totaling 35,910 pairs. Since the real-world subset is mainly constructed from driving scenes with reliable pixel-level annotations, it inevitably contains an urban-driving bias, including a relatively high proportion of transportation-related objects. As illustrated in Fig. 3, the dataset provides a diverse distribution of query semantics and weather conditions. Compared with prior reasoning segmentation benchmarks (e.g., ReasonSeg with 1.2k samples), it substantially enlarges the evaluation scale while introducing structured environmental perturbations, providing a comprehensive benchmark for assessing VLM robustness under adverse conditions.

Table 1 provides multi-dimensional comparisons between our WeatherReasonSeg and other relevant benchmarks. As shown in the table, WeatherReasonSeg stands as the only benchmark that simultaneously supports bounding boxes, pixel-level masks, general reasoning supervision, and explicit adverse-weather settings. Existing weather-oriented datasets focus on perception under degraded conditions but lack reasoning queries, while recent reasoning-centric benchmarks assume clean visual environments and do not model environmental disturbances. WeatherReasonSeg bridges this gap by unifying pixel-grounded reasoning and diverse adverse-weather scenarios within a single framework.

4 Experiments

4.1 Implementation Details

Baselines We evaluate six representative frameworks on WeatherReasonSeg, including two grounding-based methods, GLaMM [37] and Grounded-SAM [39],

Table 2: Performance on WeatherReasonSeg under different weather types and severities. Seg-Zero* is the reasoning-only model without SAM (bbox evaluation).

Method	Fog				Rain				Snow			
	val		test		val		test		val		test	
	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU
Clean												
Grounded SAM	26.0	14.5	21.3	16.4	26.0	14.5	21.3	16.4	26.0	14.5	21.3	16.4
PixelLM	26.9	22.4	14.4	13.6	26.9	22.4	14.4	13.6	26.9	22.4	14.4	13.6
GLaMM	49.1	50.9	42.7	38.9	49.1	50.9	42.7	38.9	49.1	50.9	42.7	38.9
LISA-7B	53.6	52.3	48.7	48.8	53.6	52.3	48.7	48.8	53.6	52.3	48.7	48.8
Seg-R1	60.8	56.2	55.3	46.6	60.8	56.2	55.3	46.6	60.8	56.2	55.3	46.6
Seg-Zero*	65.3	–	59.3	–	65.3	–	59.3	–	65.3	–	59.3	–
Seg-Zero-7B	62.6	62.0	57.5	52.0	62.6	62.0	57.5	52.0	62.6	62.0	57.5	52.0
Severity: Light												
Grounded SAM	23.4	12.3	18.5	14.3	22.2	13.6	16.4	15.3	22.9	13.0	16.2	15.0
PixelLM	26.2	19.6	13.4	12.1	25.3	21.8	13.7	12.2	24.4	21.4	14.2	14.1
GLaMM	46.9	48.6	41.2	37.9	46.8	49.3	41.8	38.3	44.4	46.5	41.7	38.9
LISA-7B	44.3	50.4	41.8	47.2	44.5	49.1	41.4	47.1	41.5	48.2	41.0	47.5
Seg-R1	48.4	44.3	46.9	40.8	48.5	41.5	43.9	42.5	40.1	36.9	37.7	35.8
Seg-Zero*	58.4	–	55.9	–	59.7	–	56.1	–	53.7	–	55.0	–
Seg-Zero-7B	57.1	57.4	53.9	48.4	58.6	56.8	54.0	48.6	53.2	56.8	53.1	47.0
Severity: Moderate												
Grounded SAM	19.9	14.2	16.1	13.9	17.3	12.9	14.8	13.4	16.1	12.3	13.5	12.9
PixelLM	25.7	26.2	13.6	12.0	24.2	21.8	13.6	11.4	23.8	21.0	12.3	11.6
GLaMM	46.6	47.8	40.6	38.5	46.2	46.8	41.0	37.6	42.6	46.2	38.9	37.3
LISA-7B	41.8	48.1	40.7	46.2	42.3	47.2	39.2	46.6	38.9	43.5	37.4	46.1
Seg-R1	50.8	46.6	47.8	41.9	52.9	48.1	46.6	43.5	40.8	38.6	44.2	44.3
Seg-Zero*	58.2	–	53.2	–	57.9	–	53.7	–	52.3	–	47.0	–
Seg-Zero-7B	56.0	57.3	50.9	45.3	56.6	56.1	51.5	47.0	52.0	50.3	45.7	44.5
Severity: Heavy												
Grounded SAM	15.8	11.2	12.8	12.3	14.3	10.8	12.2	11.8	13.9	11.6	11.7	10.8
PixelLM	25.0	25.9	12.2	12.7	24.0	20.3	13.2	11.3	21.9	20.0	12.7	13.3
GLaMM	46.2	45.7	39.5	38.1	44.4	45.0	39.9	37.6	42.2	45.7	38.1	36.7
LISA-7B	40.7	47.7	39.4	48.4	40.1	46.9	38.7	45.2	38.2	42.9	36.6	44.3
Seg-R1	46.4	35.5	43.4	36.5	42.3	39.3	41.5	40.3	37.1	31.1	35.4	34.9
Seg-Zero*	57.5	–	51.7	–	54.4	–	50.5	–	47.7	–	48.0	–
Seg-Zero-7B	55.4	53.4	49.9	44.0	54.3	54.3	47.8	47.7	45.9	49.7	45.7	39.7

and four reasoning-based models, PixelLM [40], LISA [22], Seg-R1 [60], and Seg-Zero [30].

Evaluation Metrics Following previous works [14, 61], we calculate gIoU and cIoU. The gIoU is the average of all per-image Intersection-over-Unions (IoUs), while the cIoU calculates the cumulative intersection over the cumulative union.

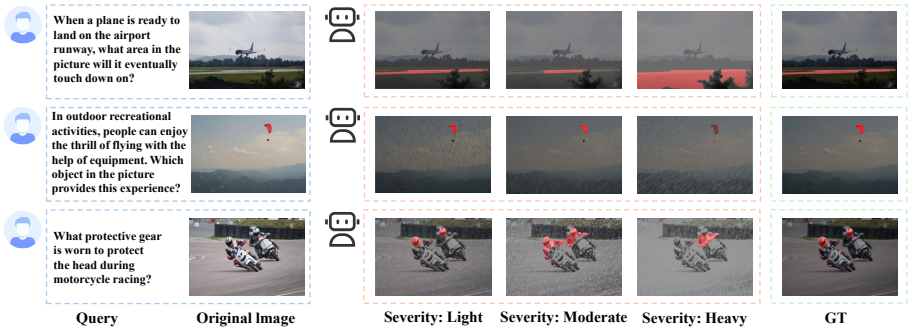

Fig. 4: A schematic comparison of VLM reasoning segmentation results under different weather degradation levels. The VLM-based segmentation method used is Seg-Zero. The figure illustrates how segmentation performance progressively changes as weather severity increases..

Unless specified, we use gIoU as our default metric, as it equally considers both large and small objects.

4.2 Results on Synthesis Segmentation Benchmarks

We first evaluate model performance on the WeatherReasonSeg-Synthesis benchmark, where weather degradations (rain, fog, and snow) are progressively introduced at three severity levels (severity: Light/Moderate/Heavy). This controlled setting enables quantitative assessment of how reasoning robustness degrades with increasing weather intensity.

As shown in Tab. 2, all models perform well under clean conditions without weather interference. However, as weather severity increases from mild (severity: Light) to severe (severity: Heavy), all methods show a clear and consistent decline in both gIoU and cIoU. This indicates that adverse weather introduces systematic disruptions to VLMs reasoning. Notably, in the Seg-Zero-7B* setting, which involves no additional segmentation optimization and thus directly reflects the VLM's reasoning quality, the gIoU decreases from 65.3 in clean conditions to 47.7 under severe weather, representing an absolute drop of 17.6. This directly shows that weather-induced visual degradation weakens the VLM's semantic association and spatial reasoning ability.

The full Seg-Zero-7B model also exhibits substantial performance drops. In foggy conditions, the gIoU decreases from 57.5 to 49.9, and in snowy scenes, the cIoU declines from 52 to 39.72. Grounded-SAM achieved a cIoU of only 10.8 in heavy fog. The impact of weather types is not uniform. Snow leads to the most severe degradation, followed by fog and then rain. This pattern is consistent with the strong loss of texture, contrast, and depth cues in snowy scenes.

Figure 4 further provides qualitative examples of Seg-Zero under increasing weather severity. Under light weather interference, the model can still produce

Table 3: Performance comparison on real-world adverse weather. Gray rows indicate model grouping.SAM2 serves as the performance upper bound because it is directly prompted with ground-truth spatial locations (e.g., bounding boxes).The overall best performances are shown in **bold**, while the second best performances are shown underlined.

Method	fog		Rain		Snow		Night	
	gIoU↑	cIoU↑	gIoU↑	cIoU↑	gIoU↑	cIoU↑	gIoU↑	cIoU↑
SAM2 (Upper Bound)								
SAM2	82.2	91.6	80.1	87.5	81.5	89.5	74.8	84.4
Reasoning and Segmentation								
PixelLM	2.8	6.0	2.0	6.9	1.9	5.2	2.6	7.6
GLaMM	13.1	14.7	9.3	9.0	11.5	12.4	9.8	11.4
Grounded SAM	23.7	14.4	18.8	16.5	15.3	14.8	11.4	12.1
LISA-7B	18.9	24.4	11.5	18.1	14.9	23.6	14.7	25.3
Seg-R1	30.9	44.4	26.7	36.1	29.4	37.9	18.5	26.3
Seg-Zero-7B	**40.6**	**47.1**	**35.1**	**38.7**	**36.8**	**41.6**	**28.9**	**29.4**

relatively accurate segmentation results. However, as weather conditions progressively deteriorate, the segmentation quality declines noticeably. In moderately degraded scenarios, foggy and snowy scenes already exhibit incomplete object localization and spatial drift. Under severe degradation, all weather types show significant segmentation errors or incomplete predictions. These observations suggest that the severe loss of visual cues caused by adverse weather significantly limits the segmentation reasoning capability of vision-language models.

4.3 Results on Real World Segmentation Benchmarks

To further evaluate VLM reasoning performance under real-world adverse weather conditions, we conduct experiments on the WeatherReasonSeg-RealWorld benchmark, which contains naturally degraded images captured under fog, rain, snow, and nighttime conditions. We first measure the upper bound of perception robustness using the pure segmentation model SAM2, and then evaluate VLM-based reasoning and segmentation methods under the same settings. As shown in Tab. 3, SAM2 achieves strong and stable performance across all conditions, with gIoU scores of 82.2 (fog), 80.1 (rain), 81.5 (snow), and 74.8 (night). Although adverse weather slightly reduces accuracy, the overall performance remains high, establishing a robust perception upper bound.

In contrast, VLM-driven reasoning segmentation methods exhibit substantial performance degradation. The best-performing Seg-Zero-7B achieves only 40.6 gIoU and 47.1 cIoU in fog, and further drops to 28.9 gIoU and 29.4 cIoU at night—approximately half of SAM2's performance under identical conditions. Grounded-SAM performs even worse, with cIoU falling below 15% at night. These results indicate a clear performance gap between perception-only and reasoning-based models in real-world degraded environments.

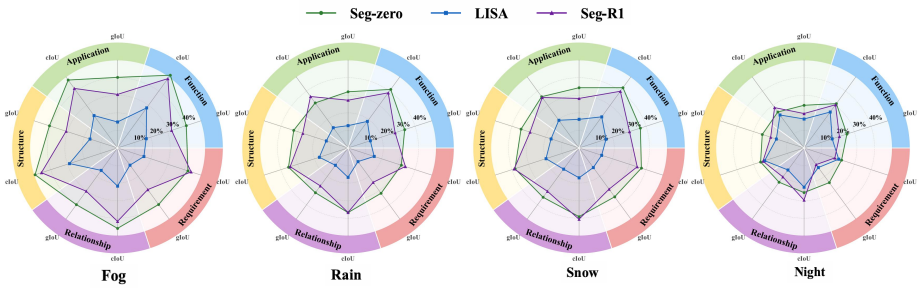

Fig. 5: Radar chart comparison of reasoning segmentation performance across five reasoning dimensions under different weather conditions (Fog, Rain, Snow, and Night). Results are reported using gIoU and cIoU for three representative VLM-based methods: Seg-Zero, LISA, and Seg-R1. The figure shows that visually grounded dimensions such as Structure and Function remain relatively stable, while context-dependent dimensions including Application and Requirement exhibit larger performance degradation under adverse weather conditions.

Next, we analyze model performance across five semantic query dimensions: Function, Application, Structure, Relationship, and Requirement. As shown in Figure 5 and summarized in Figure 4, consistent trends appear across all weather conditions. The Structure and Function dimensions consistently achieve higher segmentation performance, while Application and Requirement show noticeably lower accuracy. For example, under rainy conditions, Seg-Zero achieves a cIoU of 41.2 for Function queries but only 31.8 for Requirement queries, indicating a nearly 10% gap. This pattern suggests that queries grounded in directly observable visual attributes remain relatively robust, whereas those requiring higher-level contextual reasoning suffer greater degradation under adverse weather.

4.4 Discussions

Impact of Weather Severity on Reasoning Stability Figure 2 shows that weather severity acts as a graded stress test for VLM-driven reasoning segmentation. While all models perform strongly under clean conditions, performance consistently declines as severity increases from mild to severe across both gIoU and cIoU. This monotonic degradation indicates that adverse weather systematically disrupts pixel-grounded reasoning.

The Seg-Zero-7B* setting highlights this instability. Since this configuration reflects pure reasoning quality, the gIoU drop from 65.3 to 47.7 under severe weather directly reveals that visual degradation weakens semantic association and spatial grounding. Grounded-SAM nearly collapses under heavy fog. These results suggest that low-level visual corruption propagates to high-level reasoning, causing structural instability in the pipeline.

Impact of Real-World Adverse Weather on VLM Reasoning Even under real-world adverse conditions, the performance gap between perception-

Table 4: Reasoning performance across five query dimensions under different weather conditions. The five dimensions correspond to Function, Application, Structure, Relationship, and Requirement.The overall best performances are shown in **bold**, while the second best performances are shown underlined.

Weather	Dimension	Seg-Zero		LISA		Seg-R1	
		gIoU↑	cIoU↑	gIoU↑	cIoU↑	gIoU↑	cIoU↑
Fog	Function	**41.2**	**51.2**	**17.2**	28.2	**32.4**	**48.8**
	Application	40.2	47.8	14.8	22.8	30.5	41.9
	Structure	40.8	49.4	16.5	**28.8**	30.8	45.6
	Relationship	39.9	45.8	15.8	21.7	30.4	41.8
	Requirement	39.8	42.2	12.1	15.8	29.3	44.1
Rain	Function	33.8	41.2	13.3	18.7	28.1	38.8
	Application	32.0	31.7	12.7	14.3	27.2	36.2
	Structure	32.4	35.5	13.0	17.1	27.0	34.5
	Relationship	31.5	36.8	12.2	16.7	26.7	36.7
	Requirement	31.9	31.8	9.6	15.6	24.2	34.2
Snow	Function	36.6	42.4	16.1	21.9	30.1	39.9
	Application	34.3	36.1	16.3	19.5	28.2	35.5
	Structure	34.6	38.2	16.7	19.7	29.2	38.7
	Relationship	34.7	39.1	14.9	17.0	30.5	40.6
	Requirement	34.4	37.2	13.8	13.5	29.0	34.7
Night	Function	25.6	31.2	16.8	25.3	21.1	30.4
	Application	24.3	25.3	16.5	23.2	19.6	28.4
	Structure	24.9	25.9	16.5	23.5	19.9	24.5
	Relationship	24.1	25.4	15.5	22.3	20.4	29.7
	Requirement	24.4	22.4	13.8	20.7	11.7	18.4

only and reasoning-based models remains substantial. As shown in Figure 3, SAM2 maintains relatively strong segmentation accuracy across fog, rain, snow, and nighttime scenarios, whereas all reasoning-based methods experience significant performance degradation under the same conditions. Notably, across all weather types, reasoning and segmentation models fail to close the gap with SAM2, indicating that the primary bottleneck does not originate from low-level feature extraction. Instead, instability arises at the semantic reasoning stage, where degraded visual inputs disrupt reliable visual-semantic alignment before segmentation begins.

Moreover, the impact of different weather conditions is not uniform. Nighttime and fog lead to the most severe performance drops, while rain causes comparatively smaller degradation. This suggests that environments involving strong illumination loss or reduced global visibility more severely impair the visual cues required for reliable visual-semantic reasoning.

Performance Across Different Semantic Reasoning Dimensions Figure 4 reveals substantial differences in VLMs' reasoning performance across the five semantic query dimensions, highlighting an internal imbalance in their rea-

soning capabilities. Structure- and Function-oriented queries consistently outperform Application- and Requirement-based queries. This discrepancy arises because Application and Requirement queries rely more heavily on contextual abstraction and scenario-level reasoning, rather than directly observable visual attributes. As a result, they are more sensitive to degraded environmental cues.

Insights for Future Research Our benchmark provides several important implications for future research. First, although VLM-driven reasoning segmentation achieves strong performance under ideal conditions, VLMs exhibit significant degradation under real-world adverse weather. The primary bottleneck lies in the absence of weather-aware reasoning mechanisms capable of modeling uncertainty and adapting semantic grounding under distribution shift. Second, the observed performance differences across weather types and semantic reasoning dimensions suggest that future benchmarks should incorporate a broader range of environmental scenarios and higher-level reasoning tasks to better evaluate robustness. Finally, since current VLMs lack explicit modeling of severe weather conditions, future research could explore degradation-aware supervision, such as synthetic-to-real fine-tuning, to help VLMs extract more reliable semantic cues from visually corrupted inputs.

5 Conclusion

In this paper, we introduce WeatherReasonSeg, the first benchmark for systematically evaluating reasoning-based segmentation of vision-language models (VLMs) under adverse weather conditions. The benchmark includes two complementary components: a controllable synthetic weather dataset with multiple weather types and severity levels for robustness analysis, and a real-world adverse-weather reasoning dataset constructed via mask-guided LLM prompting to ensure semantically consistent and spatially grounded queries. To comprehensively assess reasoning behavior, we further organize the queries across five semantic dimensions: Function, Application, Structure, Relationship, and Requirement.Extensive experiments reveal the limitations of existing reasoning-based segmentation approaches and provide a comprehensive performance analysis of this underexplored problem. Our results show that current VLMs lack environment-adaptive reasoning mechanisms, leading to substantial performance degradation under adverse weather. We hope this benchmark will serve as a foundation for advancing weather-aware and robust reasoning systems in safety-critical real-world applications.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61976017), the Beijing Natural Science Foundation(4202056).

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence **39**(12), 2481–2495 (2017)
2. Bai, Z., He, T., Mei, H., Wang, P., Gao, Z., Chen, J., Zhang, Z., Shou, M.Z.: One token to seg them all: Language instructed reasoning segmentation in videos. Advances in Neural Information Processing Systems **37**, 6833–6859 (2024)
3. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
4. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
5. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
6. Chen, Y.C., Li, W.H., Sun, C., Wang, Y.C.F., Chen, C.S.: Sam4mllm: Enhance multi-modal large language model for referring expression segmentation. In: European Conference on Computer Vision. pp. 323–340. Springer (2024)
7. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
8. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. Advances in neural information processing systems **34**, 17864–17875 (2021)
9. Feng, X., Wan, Z., Wen, M., McAleer, S.M., Wen, Y., Zhang, W., Wang, J.: Alphazero-like tree-search can guide large language model decoding and training. arXiv preprint arXiv:2309.17179 (2023)
10. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
11. Hsu, J., Mao, J., Tenenbaum, J.B., Goodman, N.D., Wu, J.: What makes a maze look like a maze? International Conference on Learning Representations (ICLR) (2025)
12. Hu, X., Fu, C.W., Zhu, L., Heng, P.A.: Depth-attentional features for single-image rain removal. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 8022–8031 (2019)
13. Huang, S., Zhang, C., Ke, F., Cai, Z., Haffari, G., Qu, L., Rezatofighi, H.: Minibehavior-gran: Revealing u-shaped effects of instruction granularity on language-guided embodied agents. arXiv preprint arXiv:2604.17019 (2026)
14. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 787–798 (2014)
15. Ke, F., Cai, Z., Jahangard, S., Wang, W., Haghighi, P.D., Rezatofighi, H.: Hydra: A hyper agent for dynamic compositional visual reasoning. In: European Conference on Computer Vision. pp. 132–149. Springer (2024)
16. Ke, F., Cai, Z., Li, B., Chen, L., Lin, B., Wang, W., Haghighi, P.D., Haffari, G., Rezatofighi, H.: View2space: Studying multi-view visual reasoning from sparse observations. arXiv preprint arXiv:2603.16506 (2026)

17. Ke, F., Hsu, J., Cai, Z., Ma, Z., Zheng, X., Wu, X., Huang, S., Wang, W., Haghighi, P.D., Haffari, G., et al.: Explain before you answer: A survey on compositional visual reasoning. arXiv preprint arXiv:2508.17298 (2025)
18. Ke, F., Leng, X., Cai, Z., Khan, Z., Wang, W., Haghighi, P.D., Rezatofighi, H., Chandraker, M., et al.: Dwim: Towards tool-aware visual reasoning via discrepancy-aware workflow generation & instruct-masking tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3378–3389 (2025)
19. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
20. Kumar, A., Zhuang, V., Agarwal, R., Su, Y., Co-Reyes, J.D., Singh, A., Baumli, K., Iqbal, S., Bishop, C., Roelofs, R., et al.: Training language models to self-correct via reinforcement learning. arXiv preprint arXiv:2409.12917 (2024)
21. Lab, E.: Open R1 Multimodal. `https://github.com/EvolvingLMMs-Lab/open-r1-multimodal` (2025)
22. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9579–9589 (2024)
23. Li, B., Lin, Z., Peng, W., Nyandwi, J.d.D., Jiang, D., Ma, Z., Khanuja, S., Krishna, R., Neubig, G., Ramanan, D.: Naturalbench: Evaluating vision-language models on natural adversarial samples. Advances in Neural Information Processing Systems **37**, 17044–17068 (2024)
24. Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., Li, C.: Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. arXiv preprint arXiv:2407.07895 (2024)
25. Li, X., Juvekar, A., Zhang, J., Liu, X., Wahed, M., Nguyen, K.A., Shen, Y., Yu, T., Lourentzou, I.: Counterfactual segmentation reasoning: Diagnosing and mitigating pixel-grounding hallucination. arXiv preprint arXiv:2506.21546 (2025)
26. Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., Cobbe, K.: Let's verify step by step. In: The Twelfth International Conference on Learning Representations (2023)
27. Lin, G., Milan, A., Shen, C., Reid, I.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1925–1934 (2017)
28. Lin, J., Zhu, C., Kneuertz, P.J., Bai, Y., Xue, Y.: When models learn to ask why: Adaptive causal reasoning for trustworthy medical vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5556–5568 (2026)
29. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
30. Liu, Y., Peng, B., Zhong, Z., Yue, Z., Lu, F., Yu, B., Jia, J.: Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. arXiv preprint arXiv:2503.06520 (2025)
31. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
32. Ma, C., Jiang, Y., Wu, J., Yuan, Z., Qi, X.: Groma: Localized visual tokenization for grounding multimodal large language models. In: European Conference on Computer Vision. pp. 417–435. Springer (2024)

33. Ma, Z., Hong, J., Gul, M.O., Gandhi, M., Gao, I., Krishna, R.: Crepe: Can vision-language foundation models reason compositionally? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10910–10921 (2023)
34. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. pp. 3195–3204 (2019)
35. OpenAI: OpenAI o1. https://openai.com/o1/ (2024)
36. Qian, S., Chen, W., Bai, M., Zhou, X., Tu, Z., Li, L.E.: Affordancellm: Grounding affordance from vision language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7587–7597 (2024)
37. Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R.M., Xing, E., Yang, M.H., Khan, F.S.: Glamm: Pixel grounding large multimodal model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13009–13018 (2024)
38. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
39. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024)
40. Ren, Z., Huang, Z., Wei, Y., Zhao, Y., Fu, D., Feng, J., Jin, X.: Pixellm: Pixel reasoning with large multimodal model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26374–26383 (2024)
41. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
42. Sakaridis, C., Dai, D., Van Gool, L.: ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2021)
43. Sakaridis, C., Wang, H., Li, K., Jadon, A., Abbeloos, W., Reino, D.O., Van Gool, L., Dai, D., et al.: Acdc: The adverse conditions dataset with correspondences for robust semantic driving scene perception. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
44. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-okvqa: A benchmark for visual question answering using world knowledge. In: European conference on computer vision. pp. 146–162. Springer (2022)
45. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
46. Surís, D., Menon, S., Vondrick, C.: Vipergpt: Visual inference via python execution for reasoning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11888–11898 (2023)
47. Team, R.V.: R1-V. https://github.com/Deep-Agent/R1-V?tab=readme-ov-file (2025)
48. Tian, S., Wei, X., Zhang, S.: Neutral-reference prompting for vision-language models. arXiv preprint arXiv:2605.15615 (2026)
49. Trinh, T.H., Wu, Y., Le, Q.V., He, H., Luong, T.: Solving olympiad geometry without human demonstrations. Nature **625**(7995), 476–482 (2024)

50. Uesato, J., Kushman, N., Kumar, R., Song, F., Siegel, N., Wang, L., Creswell, A., Irving, G., Higgins, I.: Solving math word problems with process-and outcome-based feedback. arXiv preprint arXiv:2211.14275 (2022)
51. Wahed, M., Nguyen, K.A., Juvekar, A.S., Li, X., Zhou, X., Shah, V., Yu, T., Yanardag, P., Lourentzou, I.: Prima: Multi-image vision-language models for reasoning segmentation. arXiv preprint arXiv:2412.15209 (2024)
52. Wang, J., Ke, L.: Llm-seg: Bridging image segmentation and large language model reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1765–1774 (2024)
53. Wang, P., Li, L., Shao, Z., Xu, R., Dai, D., Li, Y., Chen, D., Wu, Y., Sui, Z.: Math-shepherd: Verify and reinforce llms step-by-step without human annotations. arXiv preprint arXiv:2312.08935 (2023)
54. Wang, R., Zhang, H.: Resanything: Attribute prompting for arbitrary referring segmentation. Advances in Neural Information Processing Systems **38**, 66893–66944 (2026)
55. Wang, W., Yue, T., Zhang, Y., Guo, L., He, X., Wang, X., Liu, J.: Unveiling parts beyond objects: Towards finer-granularity referring expression segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12998–13008 (2024)
56. Wu, J., Li, X., Li, X., Ding, H., Tong, Y., Tao, D.: Toward robust referring image segmentation. IEEE Transactions on Image Processing **33**, 1782–1794 (2024)
57. Wu, T.H., Biamby, G., Chan, D., Dunlap, L., Gupta, R., Wang, X., Gonzalez, J.E., Darrell, T.: See say and segment: Teaching lmms to overcome false premises. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13459–13469 (2024)
58. Xia, Z., Han, D., Han, Y., Pan, X., Song, S., Huang, G.: Gsva: Generalized segmentation via multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3858–3869 (2024)
59. Yang, S., Qu, T., Lai, X., Tian, Z., Peng, B., Liu, S., Jia, J.: Lisa++: An improved baseline for reasoning segmentation with large language model. arXiv preprint arXiv:2312.17240 (2023)
60. You, Z., Wu, Z.: Seg-r1: Segmentation can be surprisingly simple with reinforcement learning. arXiv preprint arXiv:2506.22624 (2025)
61. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 69–85. Springer (2016)
62. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6720–6731 (2019)
63. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)
64. Zhong, X., Tu, S., Ma, X., Jiang, K., Huang, W., Wang, Z.: Rainy wcity: A real rainfall dataset with diverse conditions for semantic driving scene understanding. In: Raedt, L.D. (ed.) Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22. pp. 1743–1749. International Joint Conferences on Artificial Intelligence Organization (2022)