Evaluating and Enhancing Negation Comprehension in Remote Sensing MLLMs

Haochen Han[1⋆], Jue Wang[2,1*], Alex Jinpeng Wang[1,3✉], and Fangming Liu[1✉]

[1] Peng Cheng Laboratory, Shenzhen, China
[2] Tsinghua University, Beijing, China
[3] Central South University, Changsha, China
Code and Dataset: https://hhc1997.github.io/RS-Neg-and-NeFo/

Abstract. Multimodal Large Language Models (MLLMs) have demonstrated remarkable success in various Remote Sensing (RS) tasks. However, their ability to comprehend negation remains underexplored, limiting deployment in real-world applications where models must explicitly identify what is false or absent, e.g., emergency responders need to locate non-flooded routes for evacuation. To comprehensively study this limitation, we introduce RS-Neg, the first benchmark to evaluate negation understanding across region-level to scene-level tasks. Specifically, we design an automated data generation pipeline for RS imagery, using LLMs to synthesize diverse negation queries, and introduce a dynamic visual focus module for verification. Our evaluation reveals that advanced RS MLLMs struggle with negation, exhibiting hallucinations and substantial performance degradation. To close this gap, we propose NeFo, a novel test-time adaptation method that explicitly incorporates the logical role of negation into the model optimization. Remarkably, using about 5% unlabeled test samples, NeFo significantly improves the negation understanding of models and shows strong generalization to unseen tasks.

Keywords: Multimodal · Negation Understanding · Remote Sensing

1 Introduction

Negation is a sine qua non of every linguistic system, helping humans describe their perceived world by specifying what is absent, false, or contrary to observation. Studies in cognitive science [11] reveal that negation understanding emerges before broader world knowledge develops: 18-month-olds can already interpret novel objects with negative sentences.

However, such basic reasoning ability remains underdeveloped in advanced AI systems—Multimodal Large Language Models (MLLM). Recent research [42] shows that even MLLM with 70 billion parameters suffer significant performance degradation when handling negation. This limitation becomes especially critical in RS vision-language tasks due to the inevitable demand for negation understanding. For example, models need to answer 'how many buildings are not

⋆ Equal contribution. ✉Corresponding authors.

flooded' in natural disaster monitoring, or locate 'routes without ice coverage' in transportation planning. Undoubtedly, misunderstanding negation could cause targets contrary to user intent and pose serious safety risks.

Recently, several attempts have been proposed to assess how well MLLMs handle negation. NegVQA [42] constructs two-choice questions covering different negation scenarios and image distributions. Gaslighting-Bench [43] introduces multiple-choice questions to evaluate models' resistance to misleading negation arguments while maintaining logical consistency. In addition, some studies [1, 34] design negation-enriched retrieval systems for testing embedding models like CLIP [31]. Despite the success, these benchmarks focus on general-domain vision-language tasks, and the proposed data generation pipelines do not apply to the fine-grained RS imagery [17, 18]. This raises two practical but untouched research questions: (1) How to comprehensively evaluate the negation understanding capability of MLLMs in RS scenarios? (2) How can we enhance the robustness of MLLMs when handling negation queries?

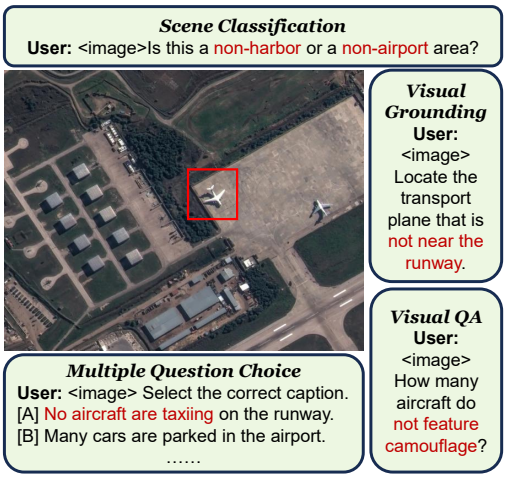


Fig. 1: A toy example to illustrate our negation understanding tasks for remote sensing.

To address the first question, we propose RS-Neg, a benchmark designed to evaluate negation understanding in RS. As illustrated in Fig. 1, RS-Neg comprises 22K samples covering four fundamental tasks that span region-level to scene-level scenarios. To construct RS-Neg, we leverage LLMs to synthesize natural and diverse negation queries from existing RS datasets, and introduce a dynamic visual focus procedure to ensure the generated negation concepts are absent from the visual content. Through our evaluation framework, we uncover that RS MLLMs face significant challenges when deployed in negation-conditioned environments.

To address the second question, we propose NeFo, an efficient Test-Time Adaptation (TTA) method that steers MLLMs to focus on negation-related semantics using only unlabeled test data. Specifically, we explicitly incorporate the logical role of negation into model optimization that maximizes the output discrepancy between the negated query and its negation-masked variant. In addition, to prevent catastrophic forgetting, we employ the original model as a teacher to regularize the predictions on affirmative queries.

The main contributions of our work are summarized as follows:

1. We introduce RS-Neg, the first benchmark for evaluating negation understanding in RS MLLMs. RS-Neg comprises 22K samples spanning region-level to scene-level scenarios, constructed through a meticulous pipeline that combines LLM-driven query synthesis with dynamic visual verification.
2. We propose NeFo, the first test-time adaptation method to enhance negation understanding in MLLMs. Without relying on additional annotations or teacher models, NeFo incorporates the logical role of negation into the optimization objective in a self-supervised manner.
3. Extensive experiments verify the effectiveness of the proposed method. With only 2.5 million training parameters and minimal unlabeled samples, NeFo achieves significant gains across different base MLLMs while demonstrating strong generalization to unseen tasks.

2 Related Works

2.1 Evaluating Negation Understanding in VLMs.

Vision-language models learn rich knowledge from large-scale multimodal data, demonstrating strong capabilities across diverse domains like RS [7, 9, 27, 41, 44]. However, recent studies have highlighted that such powerful models struggle to understand negation—a sine qua non of every human language. Based on different VLM architectures, current works can be roughly divided into two categories: 1) Embedding VLMS. To evaluate the performance of embedding models (e.g., CLIP [31] and ALIGN [14]) in negation scenarios, recent works CREPE [25] and CC-Neg [34] proposed template-based image-caption benchmarks for compositional understanding with negation. To improve the linguistic diversity of negated caption, NegBench [1] resort to LLMs to generate more natural negated captions, spanning images and videos. Evaluations on these benchmarks show that even billion-scale models like CLIP, struggles with negation in retrieval, generation, and segmentation tasks. 2) Generative VLMS. Recent studies [43] show that advanced MLLMs such as GPT-4o and Claude-3.5 can generate hallucinations when handling negation arguments from users. To evaluate this limitation in MLLMs, NegVQA [42] construct two-choice questions covering diverse negation scenarios, showing significant performance drops across 20 popular MLLMs (e.g., Qwen2-VL [36] and DeepSeek-VL [23]).

Unlike prior works that mainly provide diagnostic tools for general-domain web images, we focus on negation understanding in RS tasks. Our proposed data pipeline can handle small targets in RS images, generating comprehensive evaluation across region and scene levels.

2.2 Enhancing Negation Understanding in VLMs.

Recent methods have explored improving the negation understanding abilities in vision-language embedding models, which generally follow two research lines: 1) Post-training. Several data-driven methods attempted to fine-tune VLMs on

massive negated image-text pairs. For example, NegCLIP [39] creates negative captions via swapped relations to enhance compositional reasoning. CoNCLIP [34] integrates template-based negation samples into the contrastive loss. The recent advances use LLMs and MLLMs [1,29] to synthesize million-scale negation samples for post-training. 2) Test-time adaptation method. In contrast to data-driven methods that require massive training samples, the recent study NEAT [10] argues that the key to negation understanding is a distribution shift problem. It proposes a TTA method to correct the semantic consistency.

Despite the success, these methods target realigning vision-language embeddings and are difficult to extend to MLLMs. As collecting massive labeled RS data is expensive, we focus on TTA to update MLLMs during test time with only unlabeled test data. Given the inherent demand for negation queries in RS scenarios, we believe this study could provide some practical insights to real-world tasks such as disaster response.

3 RS-Neg: Benchmark for Negation in Remote Sensing

We propose RS-Neg, a multi-level benchmark for assessing negation comprehension in RS MLLMs, which consists of four core tasks spanning from region-level to scene-level evaluation. Specifically, for visual conversation tasks such as visual question answering, multiple choice questions, and visual grounding, we design an automated pipeline using LLMs and MLLMs to synthesize high-quality negation queries. A dynamic visual focus procedure is further introduced to localize fine-grained objects in RS images. For tasks with structured labels, e.g., scene classification, we apply rule-based reformulation to incorporate negated labels as distractors.

3.1 RS Negation Data Pipeline

Given the high cost of RS annotation, our pipeline builds upon widely used image-caption datasets, readily transforming affirmative descriptions into negation-aware counterparts. Benefiting from the powerful capabilities of LLMs or MLLMs, prior works [1, 29, 34, 42] have explored negated data augmentation through in-context learning. However, these methods focus on general-domain web images and face two key obstacles when handling RS imagery.

First, these studies mainly generate object-level negation by adding absent entities to the caption, e.g., 'a dog with no leash', while RS scenarios often demand property-level reasoning. For example, models might respond 'find a road not covered in ice' or 'how many non-flooded buildings' in disaster assistance. Second, targets of interest in RS images are typically small-scale (less than 32×32 pixels [40]) relative to the large field of view, posing significant challenges for prior pipelines with global image perception.

To address such limitations, our pipeline first uses LLMs to generate negation-conditioned captions at object, attribute, and state levels, enabling more realistic and diverse patterns. We then employ the dynamic visual focus method [16]

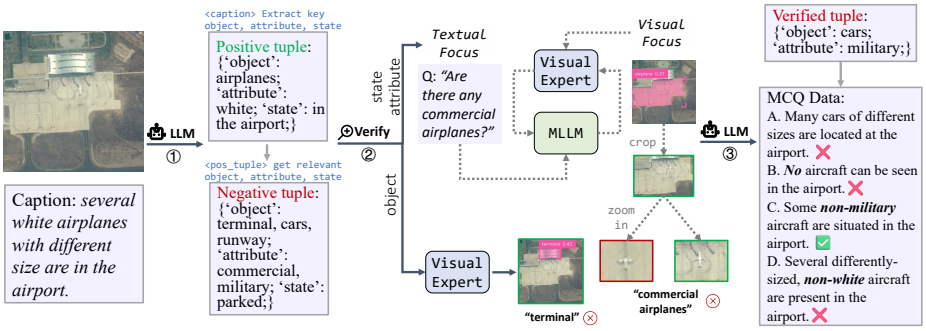


Fig. 2: Pipeline for constructing RS-Neg dataset, with MCQ as an example. Given an RS image-caption pair, we first use an LLM to extract concepts from the caption and generate corresponding negative counterparts. Second, we employ an MCTS-based visual focus method to verify these negative concepts against the image. Third, we use an LLM to formulate task-specific samples and enhance linguistic diversity.

to verify whether generated negation concepts are correct, where MLLMs and visual expert models interact via Monte Carlo Tree Search (MCTS) to focus on key visual regions. Finally, these filtered negation captions are transformed into task-specific visual conversation data via LLM augmentation. The overall pipeline proceeds as follows (Detailed steps are presented in Supplementary A.):

Extracting Candidate Negated Concepts. Given a RS image-caption pair, we first parse the caption with the LLM to extract the positive tuple, i.e., mentioned objects, corresponding attributes (e.g., color or shape), and states (e.g., action or condition). The extracted tuple is then fed to the LLM to generate the candidate negated tuple—plausible elements related to the context but absent from the caption.

Verifying Negated Concepts. Since the above process considers only text input, we then verify whether these generated negative concepts are present in the image. Although MLLMs [6,21] can accurately identify the absence of content in web images, they struggle to handle RS images due to small-scale targets. As shown in Supplementary B, even the powerful closed-source MLLM (Claude Opus 4.5) may fail in such scenarios. To address this, we adopt Dynamic Focus (DyFo) [16], a training-free MCTS-based visual search method to amplify key visual information. Specifically, for object-level elements, we directly query visual experts (e.g., Grounding Dino [22]) for presence detection. For the attribute or state element, we combine it with the corresponding object to form the root node input for tree search. In the expansion phase, new child nodes are generated by sampling from two action spaces: 1) Semantic Focus. The visual expert localizes this target and crops the specific region, yielding a smaller focused image. 2) Spatial Zoom In. As shown in Fig. 2, the cropped region may contain numerous instances of the target object (e.g., roof) with diverse properties (e.g., red and orange). We magnify the region by a scale factor, progressively narrowing down to enable instance-level inspection. The MLLM (Qwen2.5-VL-7B) then verifies

each node and provides reward signals to guide the search. Details for this process is provided in the supplementary C.

Task-Specific Data Generation. With the filtered absent elements, we formulate negation data tailored to different tasks: 1) VQA. We construct binary questions by negating either present or absent elements. For example, the target is modified as, "Does this image show {pos_object} without {neg_object}? Yes.", "Is there any {pos_object} not {pos_state} seen? No.", or "Is there a non-{neg_attribute} {pos_object} seen? Yes." All questions are further paraphrased by LLMs for linguistic diversity. 2) MCQ. For each image, we generate one correct caption and three distractors, with at least one containing negation. The MCQ task tests the model's ability to distinguish subtle differences among affirmative, negated, and hybrid descriptions, reflecting more realistic decision-making. All MCQs are also rewritten by LLMs for linguistic diversity. 3) Visual Grounding. For captions with bounding box annotations, we modify captions by negating absent elements while referring to the same target object, thus preserving the ground truth localization. 4) Scene Classification. Beyond the LLM-based pipeline above, we also introduce a simple rule-based reformulation that incorporates negated labels as distractors for the classification task. Examples of each task in RS-Neg are provided in the supplementary D.

3.2 Dataset and Evaluation

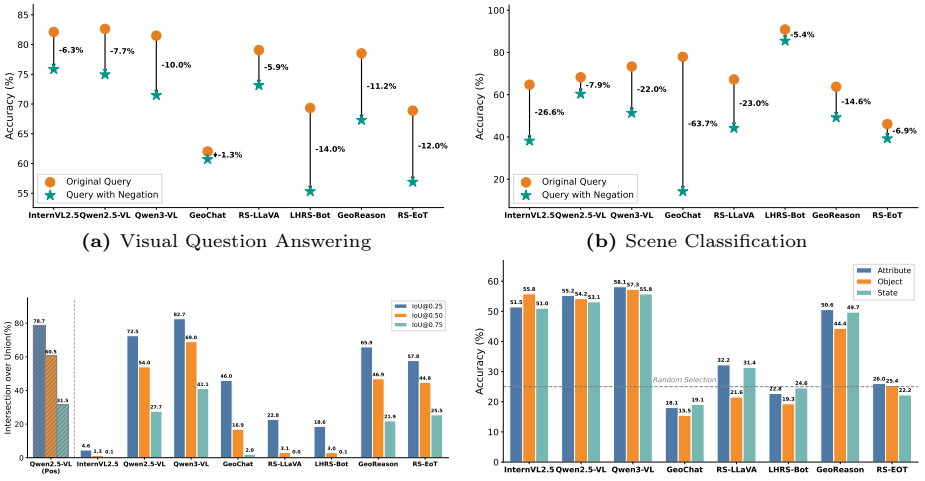
Dataset Overview. Using the proposed pipeline, we construct the RS-Neg dataset from 7 popular RS datasets, yielding 9,205 VQA samples and 5,675 MCQ samples sourced from NWPU Caption [5], RSICD [24], Sydney Caption [30], and VRSBench [20], 2,484 visual grounding samples from VRSBench [20], and 5,100 scene classification samples from AID [37] and UCMerced [38]. In total, RS-Neg contains 22,464 samples spanning four fundamental RS tasks, providing a comprehensive evaluation of negation understanding from region-level to scene-level scenarios.

Table 1: RS-Neg dataset statistics.

Task	Object	Attribute	State	Total
VQA	5,675	2,379	1,151	9,205
MCQ	4,502	791	382	5,675
Grounding	1,494	671	319	2,484
Classification	-	-	-	5,100

RS MLLMs Struggle with Negated Queries. In this section, we benchmark the negation abilities of different MLLMs using our RS-Neg, including general-purpose, RS-specific, and reasoning-augmented models. For fair comparison, we selected 8 top-performing models with similar parameter scales, i.e., Qwen2.5-VL-7B-Instruct [2], InternVL2.5-8B [4], Qwen3-VL-7B-Instruct, GeoChat-7B [15], RS-LLaVA-7B [3], LHRS-Bot-7B [26], GeoReason-7B [19], and RS-EOT-7B [33]. Note that GeoReason and RS-EOT are strong reasoning MLLMs designed to mitigate logical hallucinations in RS tasks.

Our evaluation reveals that current MLLMs consistently underperform on negation queries compared to their affirmative counterparts. As shown in Fig. 3a



(a) Visual Question Answering **(b)** Scene Classification

(c) Visual Grounding **(d)** Multiple Choice Questions

Fig. 3: Model performance on RS-Neg across different MLLMs and tasks. (a) and (b) show the performance drop compared to original queries on VQA and classification tasks, respectively. (c) reports the results on the visual grounding task, where we include the performance of Qwen2.5-VL on corresponding affirmative queries as a positive baseline. (d) reports the results on the MCQ task, where several RS-specific models exhibit notably poor performance, falling below the random guessing baseline.

and Fig. 3b, all models suffer significant performance drops on scene-level negated questions, with average degradation of 8.6% on VQA task and 21.3% on classification task. Notably, despite GeoChat's strong performance on original scene classification, its accuracy drops by 63.7% when handling negation, indicating severe overfitting to the query pattern. This limitation also holds at region-level tasks. Fig. 3c shows the visual grounding results, where we adopt the representative Qwen2.5-VL on original queries as a positive baseline given the substantial performance variance across models. Specifically, Qwen2.5-VL loses 6.2%, 6.5%, and 3.8% at IoU@0.25, IoU@0.5, and IoU@0.75, respectively. While for the more challenging MCQ task, which requires models to parse subtle yet critical differences between affirmative and negated captions, most RS-specific models (e.g., GeoChat, RS-LLaVA, and LHRS-Bot) even fall below the random guessing baseline. In addition, reasoning-augmented models such as GeoReason and RS-EOT offer no advantage, suggesting that chain-of-thought reasoning does not address the fundamental deficiency in negation understanding.

4 Method

To bridge this gap, we propose NeFo, a test-time adaptation method that guides the model to focus on negation-related semantics. The key idea behind NeFo is to explicitly incorporate the logical role [12] of negation into model optimization—a

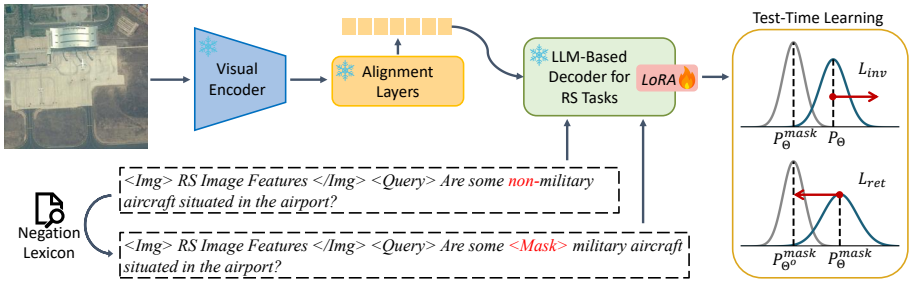


Fig. 4: An overview of the proposed NeFo. Given the test sample with a negation query, we first use a lexicon to construct the negation-masked text counterpart. Then we use LoRA to fine-tune MLLMs with two training objectives: the Truth-Value Inversion Loss that forces the model to focus on negation-related information, and the Knowledge Retaining Loss that preserves the foundational knowledge on negation-masked variants.

unary operator that inverts the truth-value of a statement. As shown in Fig. 4, without relying on additional annotations or teacher models, NeFo enhances negation understanding in a fully self-supervised manner and enables efficient adaptation through lightweight LoRA updates. We next describe the training objectives of NeFo in detail.

4.1 Problem Formulation

Without loss of generality, let $f_{\Theta^o}(\cdot)$ denote a trained Multimodal Large Language Model (MLLM) that can generate visually grounded responses for RS tasks. Although trained on massive data, text instructions are mainly expressed affirmatively, hindering model generalization for negation understanding. During real-world deployments, given a set of test samples $\{(v_i, q_i)\}_{i=1}^N$, where v_i is an RS image and q_i is the corresponding user query that may contain negation, the model can generate hallucinated outputs $f_{\Theta^o}(y|v_i, q_i)$ due to the poor ability of negation understanding.

4.2 The Training Objective

Truth-Value Inversion. From a linguistic-logical perspective [8, 12], negation acts as an operator toggling the truth value of statements between true and false. Ideally, a well-behaved MLLM should also adhere to this logical principle in its reasoning, e.g., generating totally opposite responses to 'Is the area flooded?' versus 'Is the area not flooded?'. Based on this intuition, we propose a truth-value inversion loss that maximizes the output discrepancy between the negation-conditioned query and its negation-masked variant. Specifically, given the query q containing negation, we first construct q^{mask} by masking out the negation tokens. As negation tokens in natural language are limited (e.g., no, not, and

without), we use a lexicon-based filtering as our masking strategy. Formally, let $\mathcal{P}_\Theta(y|v, q^{mask})$ denote the probability of generating output y conditioned on the affirmative input. We enforce distinct output of the negated counterpart by maximizing the KL divergence:

$$\mathcal{L}_{inv} = -\max(\mathbb{D}_{\mathrm{KL}}[\mathcal{P}_\Theta(y|v, q)\|\mathcal{P}_\Theta(y|v, q^{mask})], \alpha), \tag{1}$$

where α is a clamp parameter to ensure training stability. Equation (1) can amplify output differences arising from subtle input changes, which encourages the model to focus on key negation-related information.

Knowledge Retaining. To prevent catastrophic forgetting during TTA, we further employ the original MLLM as a teacher to regularize the model's predictions. Since the original model may produce noisy responses to negated queries, we only preserve foundational knowledge learned on negation-masked variants. Meanwhile, we introduce a standard entropy minimization as an auxiliary objective to better adapt the MLLM to the test data. Formally, the knowledge retaining loss is defined as follows:

$$\mathcal{L}_{ret} = \beta\mathbb{D}_{\mathrm{KL}}[\mathcal{P}_{\Theta^o}(y|v, q^{mask})\|\mathcal{P}_\Theta(y|v, q^{mask})] + \gamma\mathcal{H}(\mathcal{P}_\Theta(y|v, q^{mask})), \tag{2}$$

where $\mathcal{H}(\cdot)$ denote the entropy term, β and γ are balance factors.

TTA Strategy. Combining the knowledge retaining loss with truth-value inversion loss can provide clear optimization guidance for negated queries. Thus, the final TTA objective is defined as:

$$\min_{\tilde{\Theta}} \mathcal{L} = \mathcal{L}_{inv} + \mathcal{L}_{ret}, \tag{3}$$

where $\tilde{\Theta} \subseteq \Theta$ denotes the subset of model parameters to be updated. We adopt Low-Rank Adaptation to fine-tune the attention layers in the LLM module, enabling parameter-efficient training for real-world deployment.

5 Experiments

In this section, we verify the effectiveness of NeFo by addressing the following key questions: (1) Test-Time Efficacy: Does NeFo successfully enhance MLLM robustness to negation during inference? (2) Generalization Capability: Can NeFo learn generalizable negation patterns rather than simply fitting to the test data?

5.1 Implementation Details

We conduct the test-time training experiments on two negation image-caption datasets in RS-Neg, i.e., VQA and MCQ. Note that for MCQ data, we prompt each option as "Does this caption describe the image? `<image><caption>`", and only include samples with negation tokens for training. As real-world deployment requires models to adapt quickly with limited data, we perform online learning using only minimal test samples and report the performance on the full test set.

Table 2: Comparisons with state-of-the-art TTL methods on RS-Neg VQA and RS-Neg MCQ datasets under negated queries. The best results are marked in **bold**.

Method	RS-Neg VQA				RS-Neg MCQ			
	object	attribute	state	total	object	attribute	state	total
Qwen2.5-VL	80.19	69.48	60.47	74.96	54.22	55.25	53.14	54.29
·Tent	77.52	68.89	59.95	73.09	56.11	55.63	52.00	55.77
·SAR	68.35	61.75	56.91	65.21	55.84	55.37	52.09	55.52
·TLM	72.67	66.20	59.08	69.30	49.42	48.42	46.86	49.43
·NeFo	**86.26**	**72.09**	**61.69**	**79.52**	**65.73**	**65.49**	**62.30**	**65.46**
Qwen3-VL	76.00	63.64	65.33	71.47	57.29	58.15	55.76	57.30
·Tent	76.72	64.69	67.07	72.41	61.11	61.82	58.38	61.02
·SAR	79.79	65.57	67.42	74.57	61.62	62.58	59.95	61.64
·TLM	67.24	62.51	**70.03**	66.37	58.09	58.66	53.66	57.87
·NeFo	**81.04**	**66.08**	66.99	**75.42**	**64.90**	**63.84**	**61.78**	**64.55**
RS-LLaVA	77.32	65.07	69.24	73.15	21.57	32.24	31.41	23.72
·Tent	77.48	67.17	68.38	73.68	19.39	30.72	30.37	21.71
·SAR	72.09	60.99	67.33	68.63	20.19	31.23	31.41	22.48
·TLM	76.25	**66.58**	67.51	72.66	20.64	31.73	30.63	22.85
·NeFo	**79.05**	65.62	**72.46**	**74.75**	**22.41**	**32.49**	**31.94**	**24.46**
GeoReason	72.26	57.38	**63.34**	67.30	44.36	50.57	49.74	45.59
·Tent	69.96	56.07	60.34	64.88	34.12	37.42	32.72	34.48
·SAR	55.81	44.22	50.65	52.17	45.91	53.60	51.05	47.33
·TLM	65.18	**62.76**	59.25	63.81	44.85	49.56	44.76	45.50
·NeFo	**75.26**	59.77	63.25	**69.76**	**52.78**	**57.90**	**51.31**	**53.39**

Specifically, for all experiments, we use 150 and 300 samples from VQA and MCQ data for TTA, respectively. As NeFo is a general framework compatible with most MLLMs, we select four representative models on RS tasks as source models for evaluation, i.e., Qwen2.5-VL, Qwen3-VL, RS-LLaVA, and GeoReason. We use AdamW as the update with a batch size of 1. The parameters α and β are set to 0.4 and 0.4, respectively. For the LoRA setup, we set the rank to 8 and the scaling factor to 16. The learning rate and γ settings for each model are presented in supplementary E.

5.2 Comparison Experiments

In this section, we compare our method with three SOTA test-time adaptation methods, i.e., TENT [35], SAR [28], and TLM [13], under negation queries. We study both the effectiveness and generalization capability of NeFo by conducting the following analysis experiments.

Online Adaptation. Tab. 2 reports the performance on RS-Neg VQA and RS-Neg MCQ after online test-time adaptation. From the results, one could have the following observations: First, most existing TTA methods can further degrade model performance on negation understanding, which could be attributed to the self-reinforcement of wrong predictions. In contrast, NeFo could significantly improve the negation comprehension ability across diverse base models, e.g., it improves Qwen2.5-VL by 4.56% and 11.17% on RS-Neg VQA and RS-Neg MCQ, respectively. Second, across all negation types, NeFo achieves the most notable improvements on object-level queries, while gains on state-level queries are

Table 3: Zero-shot transfer evaluation on unseen datasets with negation queries. The best results are marked in **bold**.

Method	Rs-Neg	RS-Neg Grounding			FloodNet VQA		
	Classification	IoU@0.25	IoU@0.5	IoU@0.75	existence	counting	total
Qwen2.5-VL	60.37	72.54	54.03	27.66	40.14	35.23	37.32
·Tent	61.90	73.83	55.35	28.38	38.10	35.74	36.74
·SAR	61.51	70.97	53.74	27.54	31.97	36.24	34.43
·TLM	60.78	74.44	55.92	28.02	35.37	34.06	34.62
·NeFo	**66.63**	**81.16**	**63.37**	**34.46**	**82.54**	**38.93**	**57.47**
Qwen3-VL	51.31	82.65	69.00	41.10	94.78	39.26	62.87
·Tent	45.16	79.55	66.67	40.14	93.88	34.06	59.50
·SAR	48.00	83.25	69.20	**41.38**	95.47	38.76	62.87
·TLM	46.80	74.03	57.45	32.49	95.69	38.09	62.58
·NeFo	**54.75**	**83.49**	**69.24**	40.94	**96.15**	**40.10**	**63.74**
RS-LLaVA	44.18	22.75	3.06	0.04	94.10	20.13	52.36
·Tent	46.75	22.19	2.86	0.08	92.52	19.30	50.43
·SAR	31.73	20.69	2.78	0.08	93.65	20.30	51.49
·TLM	45.27	22.83	3.10	0.08	94.33	21.48	52.46
·NeFo	**48.47**	**25.89**	**4.23**	**0.28**	**95.24**	**22.32**	**53.33**
GeoReason	49.22	**65.90**	**46.94**	**21.94**	77.10	30.54	50.34
·Tent	41.59	4.31	1.97	0.60	70.75	23.49	43.59
·SAR	39.98	13.37	6.48	2.05	78.46	28.19	49.57
·TLM	50.20	27.90	16.63	5.43	78.46	28.02	49.47
·NeFo	**51.33**	64.69	45.85	**21.94**	**79.60**	**34.40**	**53.62**

relatively modest, particularly on the VQA task (average 2.46% versus 1.50%). This discrepancy likely arises because object-level negation involves concrete, easily identifiable entities, whereas state-level negation requires reasoning about context-dependent conditions, which is inherently more challenging.

Zero-shot Transferability. To investigate whether NeFo can learn generalizable negation patterns, we evaluate the TTA-updated models on unseen negation data tasks. Specifically, we use the models after TTA on RS-Neg MCQ and test them on RS-Neg Classification, RS-Neg Visual Grounding, and FloodNet VQA [32]. Note that FloodNet is a flood disaster dataset, from which we curate 1,037 VQA samples with real-world negation queries, spanning existence and counting categories. As shown in Tab. 3, for the scene-level classification task, NeFo consistently surpasses the original MLLMs, with improvements ranging from 2.11% to 6.26%. For the region-level grounding task, NeFo shows varied performance based on the original model. For example, it improves Qwen2.5-VL by 8.62%, 9.34%, and 6.80% at IoU@0.25, IoU@0.5, and IoU@0.75, respectively. Notably, this performance even surpasses the results on corresponding affirmative queries (shown in Fig. 3c), indicating that proper negation understanding provides additional discriminative information that helps the model better locate the target. In contrast, NeFo shows marginal performance drops on GeoReason (-1.21% and -1.09% at IoU@0.25 and IoU@0.5, respectively), but still far outperforms other TTA methods, e.g., TENT causes a catastrophic 61.59% drop at IoU@0.25. This may be attributed to the reasoning-augmented nature of GeoReason, where answer-level supervision of TTA could disrupt the intermediate

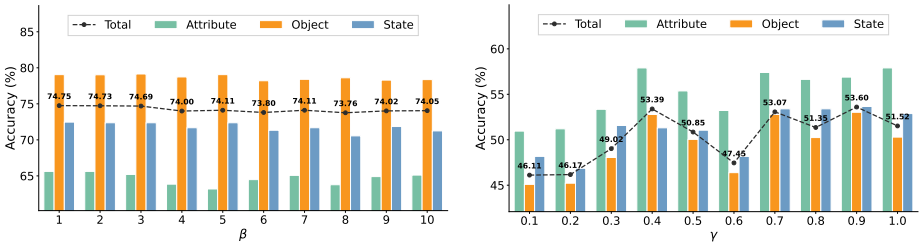
Table 4: Ablation study of each training objective on RS-Neg VQA and RS-Neg MCQ.

Method	RS-Neg VQA				RS-Neg MCQ			
	object	attribute	state	total	object	attribute	state	total
Base	80.19	69.48	60.47	74.96	54.22	55.25	53.14	54.29
·NeFo w/o TI	79.22	68.89	59.17	74.05	57.97	58.28	54.45	57.78
·NeFo w/o EM	82.71	70.24	61.51	76.84	36.54	27.23	35.08	29.06
·NeFo w/o KR	83.81	**72.68**	58.91	77.82	20.32	31.48	32.72	22.71
·NeFo	**86.26**	72.09	**61.69**	**79.52**	**65.73**	**65.49**	**62.30**	**65.46**

reasoning process. Our NeFo could mitigate this degradation through the knowledge retaining objective. For the real-world negation questions, i.e., FloodNet VQA, NeFo consistently achieves performance improvements, with an average gain of 2.63% on the challenging counting task. These zero-shot results demonstrate that NeFo enables models to acquire generalizable negation understanding rather than simply fitting to the test data.

5.3 Ablation Study

Impact of Each Component. To study the influence of specific components in our method, we carry out the ablation study on the RS-Neg VQA and RS-Neg MCQ using Qwen2.5-VL-7B. Specifically, we ablate the contributions of three key training objectives of NeFo, i.e., Truth-Value Inversion (TI), Knowledge Retaining (KR), and Entropy Minimization (EM). From the results in Tab. 4, we observe the following conclusions: First, the full NeFo could achieve the best performance, showing that all these objectives are important to improve the robustness against negation. Second, different tasks exhibit varying sensitivity to each component. For VQA, Truth-Value Inversion is most critical, whereas Knowledge Retaining plays a more important role in MCQ. This is likely because MCQ is more complex, containing both affirmative and negated captions, requiring the model to preserve its capability on affirmative queries to make correct distinctions.



(a) Effect of β on RS-Neg VQA **(b)** Effect of γ on RS-Neg MCQ

Fig. 5: Sensitivity analysis of hyper-parameters in NeFo. (a) RS-Neg VQA with RS-LLaVA as the base model. (b) RS-Neg MCQ with GeoReason as the base model.

Table 5: Comparisons of NeFo across different model sizes on RS-Neg VQA and RS-Neg MCQ. The best results are marked in **bold**.

Method	RS-Neg VQA				RS-Neg MCQ			
	object	attribute	state	total	object	attribute	state	total
Qwen3-VL-2B	70.29	55.82	54.74	64.61	34.74	42.48	40.31	36.19
·NeFo	**82.06**	**62.42**	**62.29**	**74.51**	**37.69**	**45.39**	**42.67**	**39.10**
Qwen3-VL-4B	84.19	66.16	64.38	77.06	61.88	62.96	57.07	61.71
·NeFo	**84.32**	**67.72**	**64.81**	**77.59**	**65.55**	**65.87**	**58.64**	**65.13**
Qwen3-VL-8B	76.00	63.64	65.33	71.47	57.29	58.15	55.76	57.30
·NeFo	**81.04**	**66.08**	**66.99**	**75.42**	**64.90**	**63.84**	**61.78**	**64.55**
Qwen3-VL-32B	84.42	66.83	66.99	77.70	64.73	65.74	59.69	64.53
·NeFo	**84.69**	**67.13**	**67.07**	**77.95**	**65.79**	**67.00**	**61.26**	**65.66**

Parameter Analysis. We now investigate the effect of hyper-parameters β and γ by varying their values and plotting the resulting accuracy. Specifically, we report the accuracy on RS-Neg VQA with RS-LLaVA for β, and on RS-Neg MCQ with GeoReason for γ. As shown in Fig. 5, we first observe that NeFo is robust to the choice of β, with VQA accuracy varying within a narrow range, i.e., 73.76% $\sim$ 74.75%. This is because a small β is sufficient to prevent knowledge forgetting—the knowledge retaining loss consistently remains near zero during TTA, making further increases in β unnecessary. In contrast, NeFo is more sensitive to γ, with performance dropping by 7.28% when γ is set to a small value ($\gamma = 0.1$). This suggests that applying entropy minimization to negation-masked captions is necessary for the challenging MCQ task. On the one hand, it helps the model adapt to the domain distribution of test data. On the other hand, since $\mathcal{L}_{inv}$ relies on affirmative predictions as anchors, strengthening the model's confidence on affirmative inputs provides a more stable optimization direction for negation understanding. In addition, when γ increases, NeFo can maintain stable performance across a broader range of values, i.e., 0.7 $\sim$ 1.0.

5.4 Scaling Analysis

Scaling Across Model Sizes. In this section, we explore the effectiveness of NeFo when applied to MLLMs of varying sizes. Specifically, we conduct experiments on the Qwen3-VL family across four model scales, i.e., 2B, 4B, 8B, and 32B. Tab. 5 presents the performance on RS-Neg VQA and RS-Neg MCQ. From the results, we could have the following observation: First, scaling up model size does not always improve performance, e.g., Qwen3-VL-4B significantly outperforms Qwen3-VL-8B on both VQA and MCQ tasks. Second, NeFo consistently improves performance across all four model scales. Notably, for the smallest 2B model, NeFo achieves a 9.9% absolute gain on RS-Neg VQA, bringing its negation understanding capability close to that of the 32B model. This highlights the practical value of NeFo for resource-constrained deployments, such as edge devices or UAV-based RS applications.

Scaling Across Data Sizes. In this section, we study the effect of training data size on NeFo by varying the number of samples used for TTA. All exper-

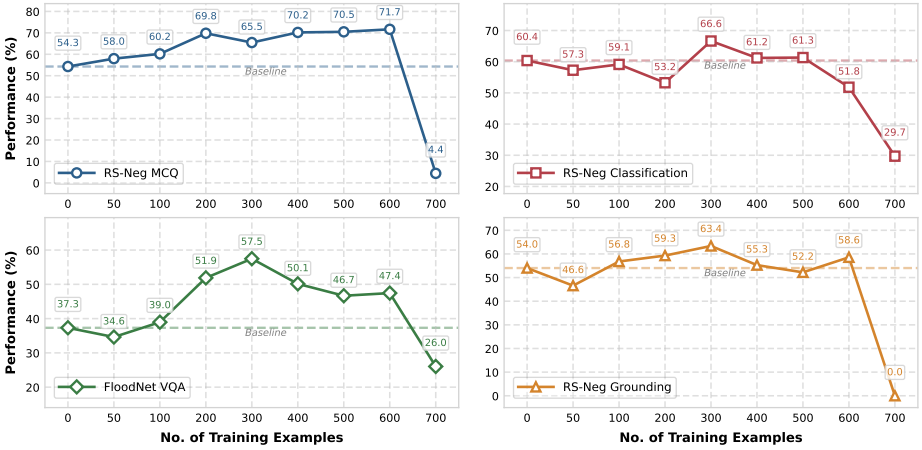


Fig. 6: Effect of NeFo with varying test-time adaptation data sizes. We report IoU@0.5 for RS-Neg Grounding and accuracy for the remaining tasks.

iments follow the same training settings for a fair comparison. We increase the training data for Qwen2.5-VL-7B on RS-Neg MCQ from 50 to 700 samples and evaluate its zero-shot performance on RS-Neg Classification, RS-Neg Grounding, and FloodNet VQA. As shown in Fig. 6, one could see that model performance on MCQ improves significantly as training data increases, achieving a 17.4% absolute gain with 600 training samples. However, further increasing the training data leads to overfitting and performance collapse, which may require a smaller learning rate to mitigate this issue. Meanwhile, the zero-shot performance on unseen datasets reveals a trade-off: insufficient training samples cause the model to only fit the negation data without learning generalizable patterns, while excessive samples may lead to overfitting and degraded generalization.

6 Conclusion

This paper studies a practical but overlooked problem in RS MLLMs: negation understanding. We introduce RS-Neg, a comprehensive benchmark covering four basic RS tasks, and reveal that current MLLMs suffer significant performance degradation under negation queries. To address this, we propose NeFo, a test-time adaptation method that incorporates negation logic into optimization via self-supervision. Experiments demonstrate that NeFo significantly improves negation understanding across different base models and generalizes well to unseen tasks. We hope this work could enhance the practicality of MLLMs in RS applications such as disaster response and emergency monitoring.

Acknowledgmetns

This work was supported in part by the Major Key Project of PCL under Grant PCL2025A10 and PCL2024A06, and in part by the Shenzhen Science and Technology Program under Grant RCJC20231211085918010.

References

1. Alhamoud, K., Alshammari, S., Tian, Y., Li, G., Torr, P.H., Kim, Y., Ghassemi, M.: Vision-language models do not understand negation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29612–29622 (2025)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Bazi, Y., Bashmal, L., Al Rahhal, M.M., Ricci, R., Melgani, F.: Rs-llava: A large vision-language model for joint captioning and question answering in remote sensing imagery. Remote Sensing **16**(9), 1477 (2024)
4. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
5. Cheng, Q., Huang, H., Xu, Y., Zhou, Y., Li, H., Wang, Z.: Nwpu-captions dataset and mlca-net for remote sensing image captioning. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–19 (2022)
6. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems **36**, 49250–49267 (2023)
7. Danish, M., Munir, M.A., Shah, S.R.A., Kuckreja, K., Khan, F.S., Fraccaro, P., Lacoste, A., Khan, S.: Geobench-vlm: Benchmarking vision-language models for geospatial tasks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7132–7142 (2025)
8. Déprez, V., Déprez, V.M., Espinal, M.T., i Farré, M.T.E.: The Oxford handbook of negation. Oxford University Press (2020)
9. Guo, X., Lao, J., Dang, B., Zhang, Y., Yu, L., Ru, L., Zhong, L., Huang, Z., Wu, K., Hu, D., et al.: Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27672–27683 (2024)
10. Han, H., Wang, A.J., Liu, F., Zhu, J.: Negation-aware test-time adaptation for vision-language models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2026)
11. Hasson, U., Glucksberg, S.: Does understanding negation entail affirmation?: An examination of negated metaphors. Journal of Pragmatics **38**(7), 1015–1032 (2006)
12. Horn, L.: A Natural History of Negation. David Hume series, CSLI (2001), `https://books.google.com.sg/books?id=hBFtAAAAIAAJ`
13. Hu, J., Zhang, Z., Chen, G., Wen, X., Shuai, C., Luo, W., Xiao, B., Li, Y., Tan, M.: Test-time learning for large language models. In: Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., Zhu, J. (eds.)

Proceedings of the 42nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 267, pp. 24823–24849. PMLR (13–19 Jul 2025)
14. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
15. Kuckreja, K., Danish, M.S., Naseer, M., Das, A., Khan, S., Khan, F.S.: Geochat: Grounded large vision-language model for remote sensing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27831–27840 (2024)
16. Li, G., Xu, J., Zhao, Y., Peng, Y.: Dyfo: A training-free dynamic focus visual search for enhancing lmms in fine-grained visual understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9098–9108 (2025)
17. Li, Q., He, X., Shu, X., Yu, Y., Chen, D., Chen, Y., Yang, X.: A simple aerial detection baseline of multimodal language models. In: IGARSS 2025-2025 IEEE International Geoscience and Remote Sensing Symposium. pp. 6833–6837. IEEE (2025)
18. Li, Q., Ma, S., Luo, J., Yu, Y., Zhou, Y., Wang, F., Lu, X., Wang, X., He, X., Chen, Y., et al.: Co-training vision-language models for remote sensing multi-task learning. Remote Sensing **18**(2), 222 (2026)
19. Li, W., Xiang, X., Wen, Z., Zhou, G., Niu, B., Wang, F., Huang, L., Wang, Q., Hu, Y.: Georeason: Aligning thinking and answering in remote sensing vision-language models via logical consistency reinforcement learning. arXiv preprint arXiv:2601.04118 (2026)
20. Li, X., Ding, J., Elhoseiny, M.: Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. Advances in Neural Information Processing Systems **37**, 3229–3242 (2024)
21. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
22. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55. Springer (2024)
23. Lu, H., Liu, W., Zhang, B., Wang, B., Dong, K., Liu, B., Sun, J., Ren, T., Li, Z., Yang, H., et al.: Deepseek-vl: towards real-world vision-language understanding. arXiv preprint arXiv:2403.05525 (2024)
24. Lu, X., Wang, B., Zheng, X., Li, X.: Exploring models and data for remote sensing image caption generation. IEEE Transactions on Geoscience and Remote Sensing **56**(4), 2183–2195 (2017)
25. 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)
26. Muhtar, D., Li, Z., Gu, F., Zhang, X., Xiao, P.: Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. In: European Conference on Computer Vision. pp. 440–457. Springer (2024)
27. Nedungadi, V., Kariryaa, A., Oehmcke, S., Belongie, S., Igel, C., Lang, N.: Mmearth: Exploring multi-modal pretext tasks for geospatial representation learning. In: European Conference on Computer Vision. pp. 164–182. Springer (2024)

28. Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., Tan, M.: Towards stable test-time adaptation in dynamic wild world. In: The Eleventh International Conference on Learning Representations (2023)
29. Park, J., Lee, J., Song, J., Yu, S., Jung, D., Yoon, S.: Know" no"better: A data-driven approach for enhancing negation awareness in clip. arXiv preprint arXiv:2501.10913 (2025)
30. Qu, B., Li, X., Tao, D., Lu, X.: Deep semantic understanding of high resolution remote sensing image. In: 2016 International conference on computer, information and telecommunication systems (Cits). pp. 1–5. IEEE (2016)
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
32. Rahnemoonfar, M., Chowdhury, T., Sarkar, A., Varshney, D., Yari, M., Murphy, R.R.: Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. IEEE Access **9**, 89644–89654 (2021)
33. Shao, R., Li, Z., Zhang, Z., Xu, L., He, X., Yuan, H., He, B., Dai, Y., Yan, Y., Chen, Y., et al.: Asking like socrates: Socrates helps vlms understand remote sensing images. arXiv preprint arXiv:2511.22396 (2025)
34. Singh, J., Shrivastava, I., Vatsa, M., Singh, R., Bharati, A.: Learn" no" to say" yes" better: Improving vision-language models via negations. arXiv preprint arXiv:2403.20312 (2024)
35. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. arXiv preprint arXiv:2006.10726 (2020)
36. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
37. Xia, G.S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., Lu, X.: Aid: A benchmark data set for performance evaluation of aerial scene classification. IEEE Transactions on Geoscience and Remote Sensing **55**(7), 3965–3981 (2017)
38. Yang, Y., Newsam, S.: Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems. pp. 270–279 (2010)
39. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? arXiv preprint arXiv:2210.01936 (2022)
40. Zhang, Y., Ye, M., Zhu, G., Liu, Y., Guo, P., Yan, J.: Ffca-yolo for small object detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–15 (2024). https://doi.org/10.1109/TGRS.2024.3363057
41. Zhang, Y., Ru, L., Wu, K., Yu, L., Liang, L., Li, Y., Chen, J.: Skysense v2: A unified foundation model for multi-modal remote sensing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9136–9146 (2025)
42. Zhang, Y., Su, Y., Liu, Y., Yeung-Levy, S.: Negvqa: Can vision language models understand negation? arXiv preprint arXiv:2505.22946 (2025)
43. Zhu, B., Qi, H., Gui, Y., Chen, J., Ngo, C.W., Lim, E.P.: Calling a spade a heart: Gaslighting multimodal large language models via negation. arXiv preprint arXiv:2501.19017 (2025)
44. Zhu, Q., Lao, J., Ji, D., Luo, J., Wu, K., Zhang, Y., Ru, L., Wang, J., Chen, J., Yang, M., et al.: Skysense-o: Towards open-world remote sensing interpretation with vision-centric visual-language modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14733–14744 (2025)